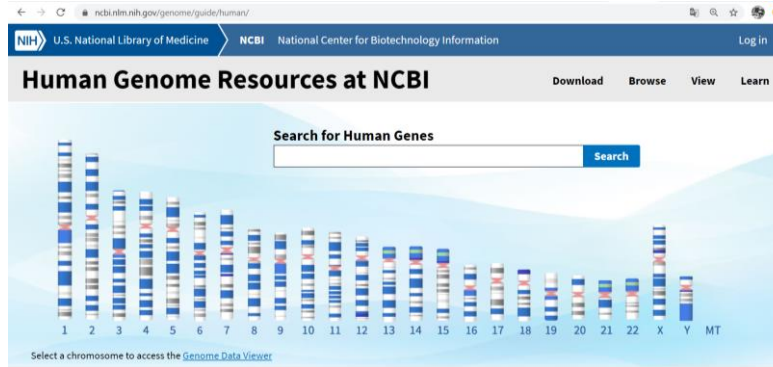
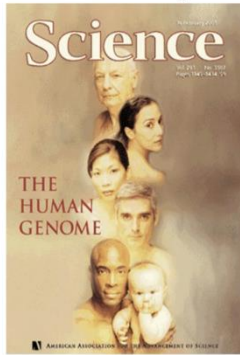


Capitolo 8

Genomica: La mappatura e il sequenziamento dei genomi

Genomica strutturale: Mappa fisica del genoma

February 2001 - Publication of the first draft of the human genome



ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=7157

Genome Data Viewer

Homo sapiens (human)

Assembly: GRCh38.p13

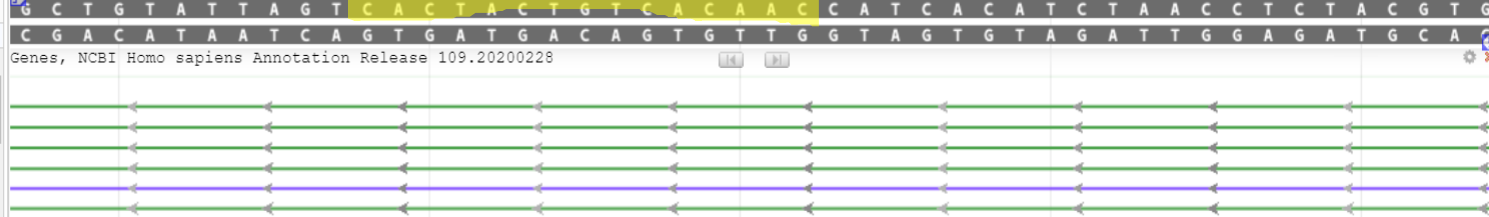
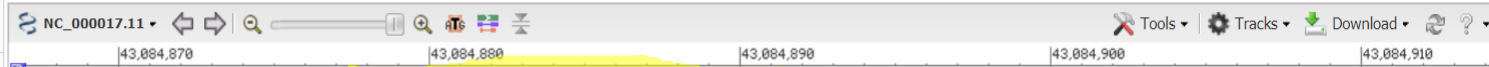
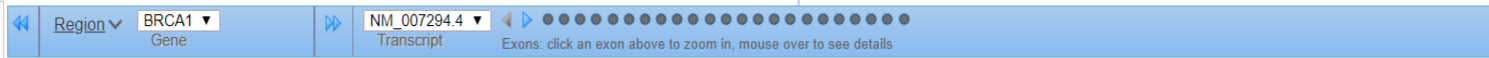
Search assembly
BRCA1

Name	Location
BRCA1	Chr17: 43,044,295 - 43,125,483
BRCA2	Chr13: 32,315,480 - 32,399,672
RAD51	Chr15: 40,694,774 - 40,732,340
BARD1	Chr2: 214,725,645 - 214,809,711
BABAM1	Chr19: 17,267,376 - 17,279,353
BAP1	Chr3: 52,401,004 - 52,410,105
BRIP1	Chr17: 61,679,186 - 61,864,120
LOC111589215	Chr17: 43,124,495 - 43,127,162

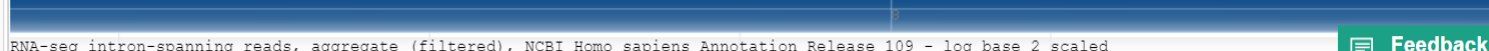
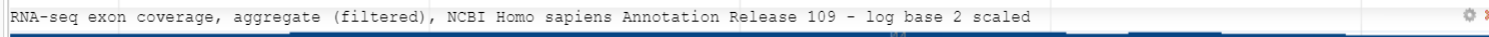
Pick Assembly

Assembly: GCF_000001405.39 (GRCh38.p13)

Sequence	Location
NC_000017.11	7,668,402 - 7,687,550



Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109...
Warning: No track data found in this range



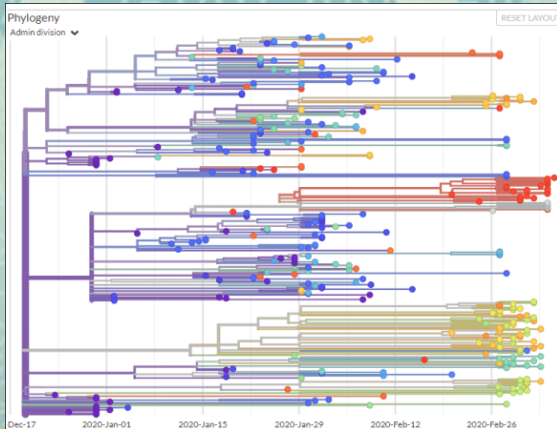
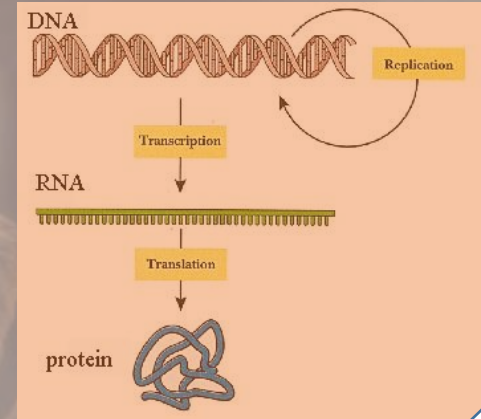
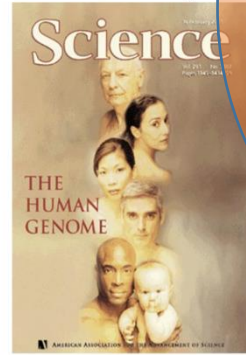
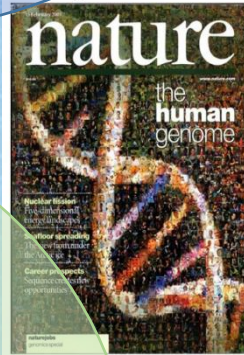
Feedback

Genomica strutturale

Genomica comparativa

Genomica funzionale

February 2001 - Publication of the first draft of the human genome





1990. 3 miliardi di dollari in 15 anni. HUGO (Human Genome Organization)

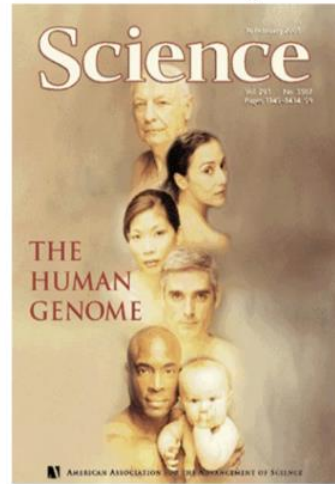


CELERA

A PECO Corporation Business

Iniziativa privata

February 2001 - Publication of the first draft of the human genome



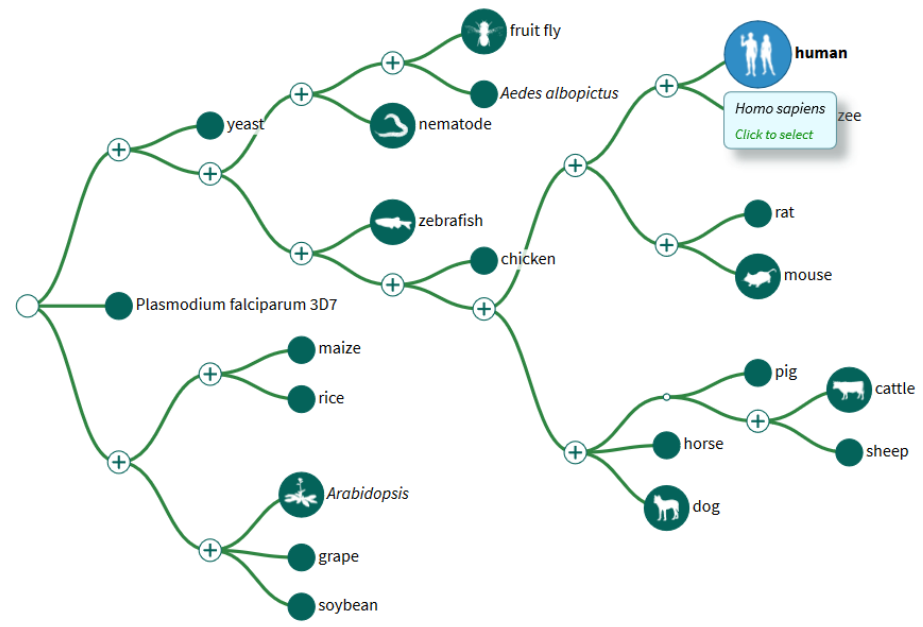
- *Escherichia coli* (in rappresentazione dei procarioti)
- Il lievito *Sccharomyces cerevisiae* (eucarioti unicellulari)
- *Drosophila melanogaster* e *Caenorhabditis elegans* (animali multicellulari con genomi moderatamente complessi)
- *Mus musculus* (mammifero).

Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 830 eukaryotic RefSeq genome assemblies.

Select organism

Homo sapiens (human)



Homo sapiens (human) genome

Search in genome
Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

Assembly
GRCh38.p13

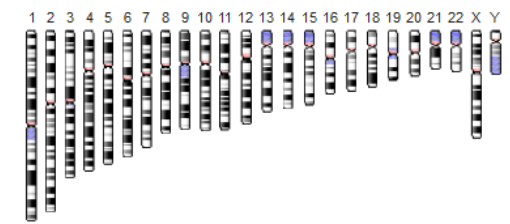
[Browse genome](#) [BLAST genome](#)

Assembly details

Name	GRCh38.p13
RefSeq accession	GCF_000001405.39
GenBank accession	GCA_000001405.28
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome

Annotation details

Annotation Release 109
Release date 2020-03-02



Procedura per il sequenziamento genomico

1. Isolamento del DNA genomico



2. Clonaggio del DNA

2.1. Frammentazione del DNA: **digestione enzimatica**

2.2. Unione del frammento di DNA a un **vettore di clonaggio**: DNA ricombinante

2.3. Introduzione del vettore ricombinante nella cellula ospite: **trasformazione**

2.4. **Coltura** delle cellule trasformate: **cloni batterici**

3. Sequenziamento del DNA: Tecnologie **Sanger** e **NGS**

Clonaggio del DNA

1. Gli enzimi di restrizione (o **endonucleasi di restrizione**). Frammentazione del DNA

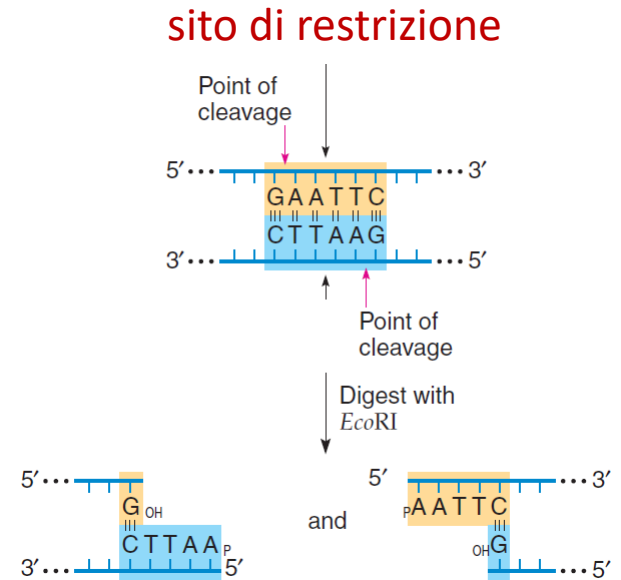
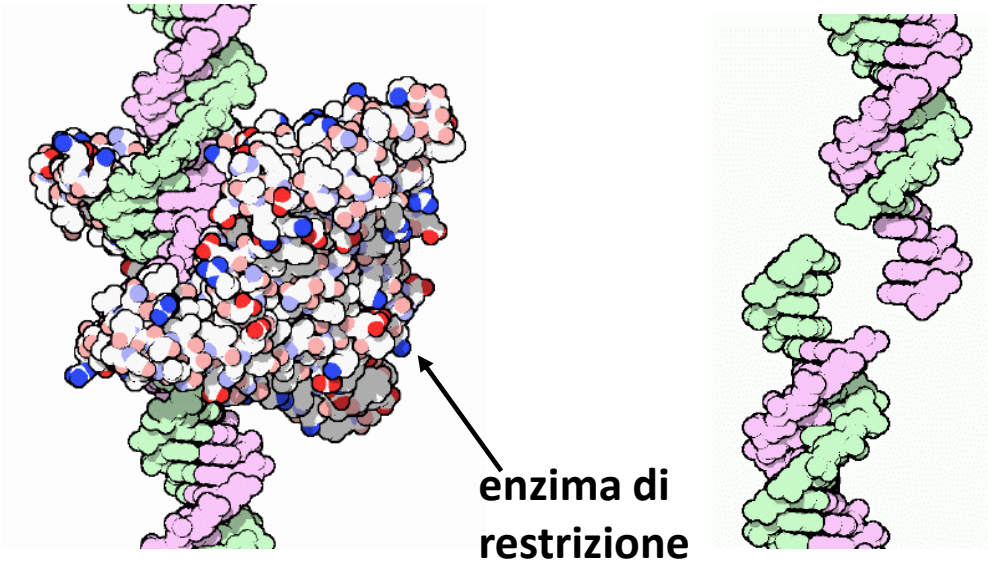
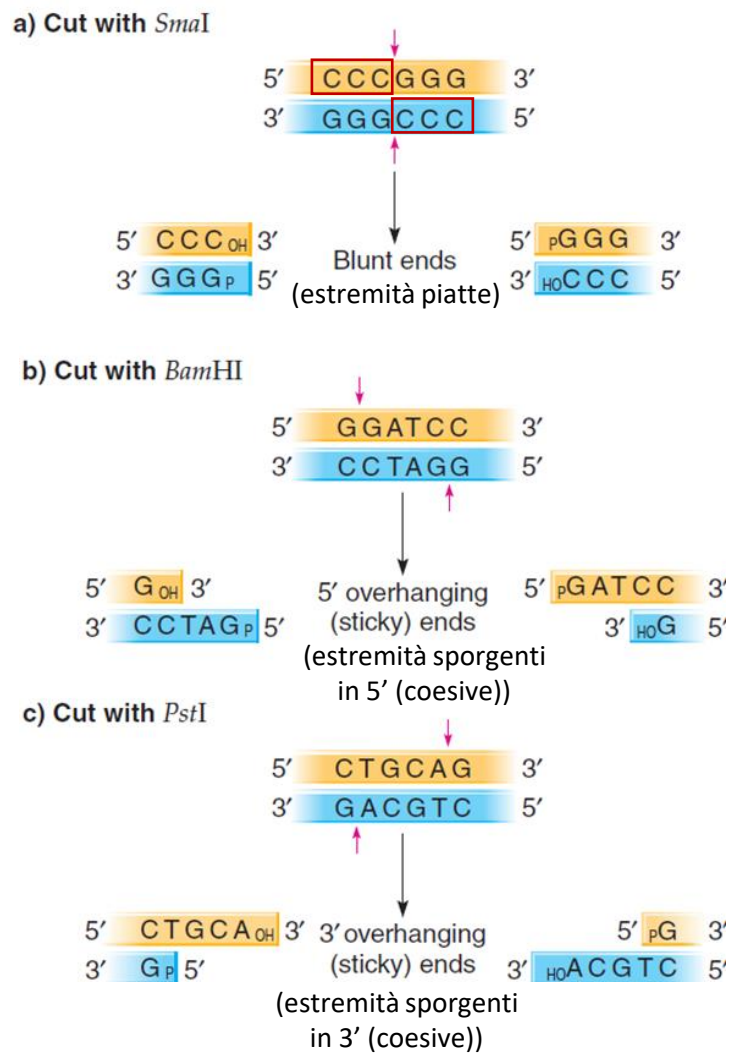


Table 8.1 Characteristics of Some Restriction Enzymes

	Enzyme Name	Pronunciation	Organism in Which Enzyme Is Found	Recognition Sequence and Position of Cut ^a
Enzymes with 6-bp Recognition Sequences	BamHI	"bam-H-one"	<i>Bacillus amyloliquefaciens</i> H	5'-G [↓] GATC C-3' 3'-CCTAG [↑] G-5'
	BglII	"bagel-two"	<i>Bacillus globigi</i>	5'-A [↓] GATCT-3' 3'-TCTAG [↑] A-5'
	EcoRI	"echo-R-one"	<i>Escherichia coli</i> RY13	5'-G [↓] AATTC-3' 3'-CTTAA [↑] G-5'
	HaeII	"hay-two"	<i>Haemophilus aegypticus</i>	5'-R [↓] GC [↓] GY-3' 3'-Y [↑] CG [↑] GR-5'
	HindIII	"hin-D-three"	<i>Haemophilus influenzae</i> R ₄	5'-A [↓] AGCTT-3' 3'-TTCGAA [↑] -5'
	PstI	"P-S-T-one"	<i>Providencia stuartii</i>	5'-C [↓] TGCA [↑] G-3' 3'-G [↑] ACGTC-5'
	SalI	"sal-one"	<i>Streptomyces albus</i>	5'-G [↓] TCTGA C-3' 3'-CAGCT [↑] G-5'
Enzymes with 4-bp Recognition Sequences	HaeIII	"hay-three"	<i>Haemophilus aegypticus</i>	5'-G [↓] GC C-3' 3'-C [↑] CG G-5'
	HhaI	"ha-ha-one"	<i>Haemophilus haemolyticus</i>	5'-G [↓] CG [↓] C-3' 3'-C [↑] G [↑] CG-5'
	HpaII	"hepa-two"	<i>Haemophilus parainfluenzae</i>	5'-C [↓] GG C-3' 3'-G [↑] GC [↑] C-5'
	Sau3A	"sow-three-A"	<i>Staphylococcus aureus</i> 3A	5'-G [↓] ATC -3' 3'-CTA G [↑] -5'
Enzyme with 8-bp Recognition Sequences	NotI	"not-one"	<i>Nocardia otitidis-caviarum</i>	5'-G [↓] C [↓] G [↓] CC [↓] GC-3' 3'-C [↑] GC [↑] CC [↑] GC-5'
Enzyme with Recognition Sequence Containing a Nonspecific Spacer Sequence	BstXI	"b-s-t-x-one"	<i>Bacillus stearothermophilus</i>	5'-CCANNNNNNTGG-3' 3'-GGTNNNNNACC-5'

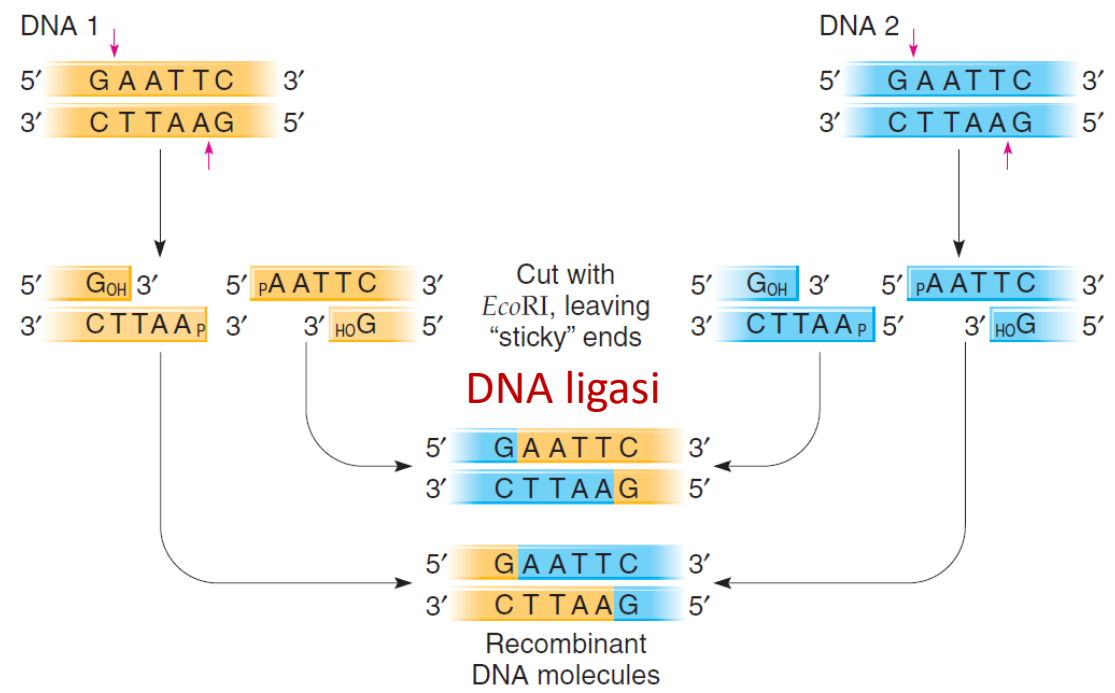
genera
specie
cepo
EcoRI

Figure 8.2 Examples of how restriction enzymes cleave DNA.



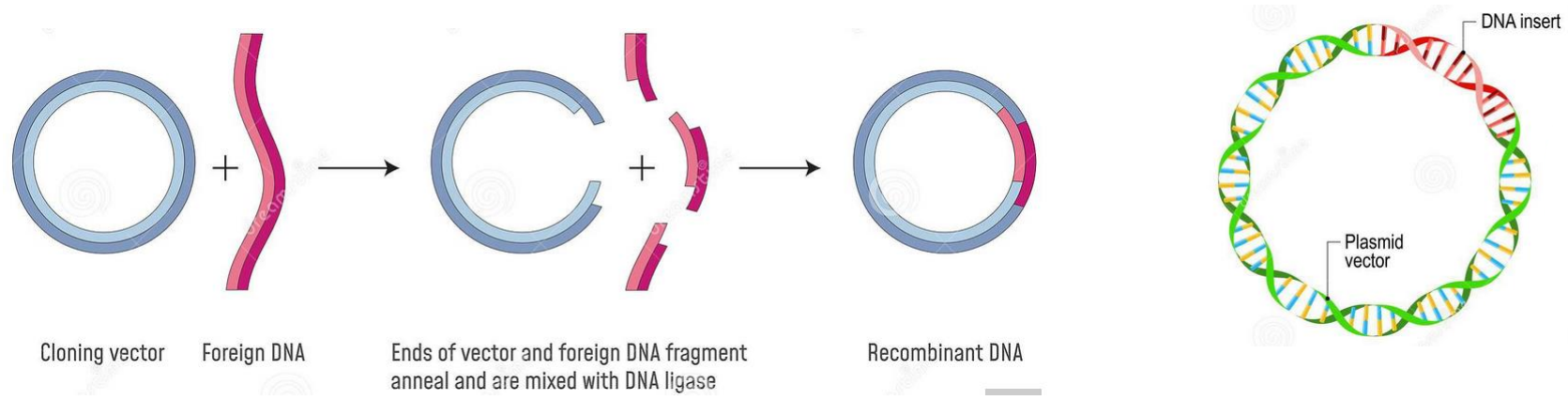
Frequenza dei siti di restrizione nel DNA. In generale, quanto più corta sia la sequenza del sito di restrizione di un enzima, più è probabile trovarlo nel genoma e più frammentato risulterà il genoma dopo della digestione enzimatica.

Gli enzimi di restrizione che producono estremità coesive sono particolarmente utili nel clonaggio del DNA



L'unione tra due frammenti di DNA di origine diversa è alla base della **tecnologia del DNA ricombinante**.

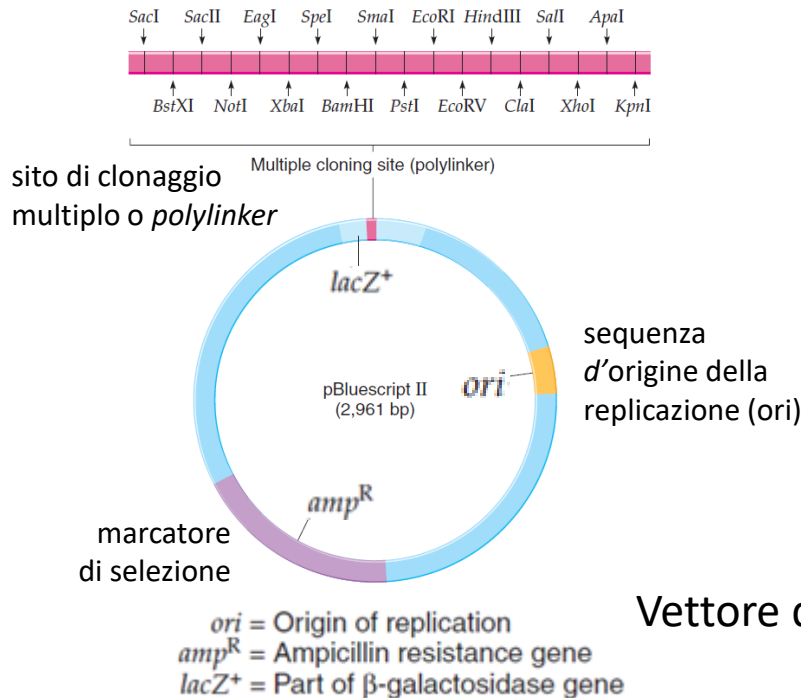
Se un vettore di clonaggio e un frammento di DNA vengono tagliati con lo stesso enzima di restrizione (per esempio *EcoRI*), avrà luogo l'appaiamento (*annealing*) delle basi nelle estremità coesive.



2. I vettori di clonaggio

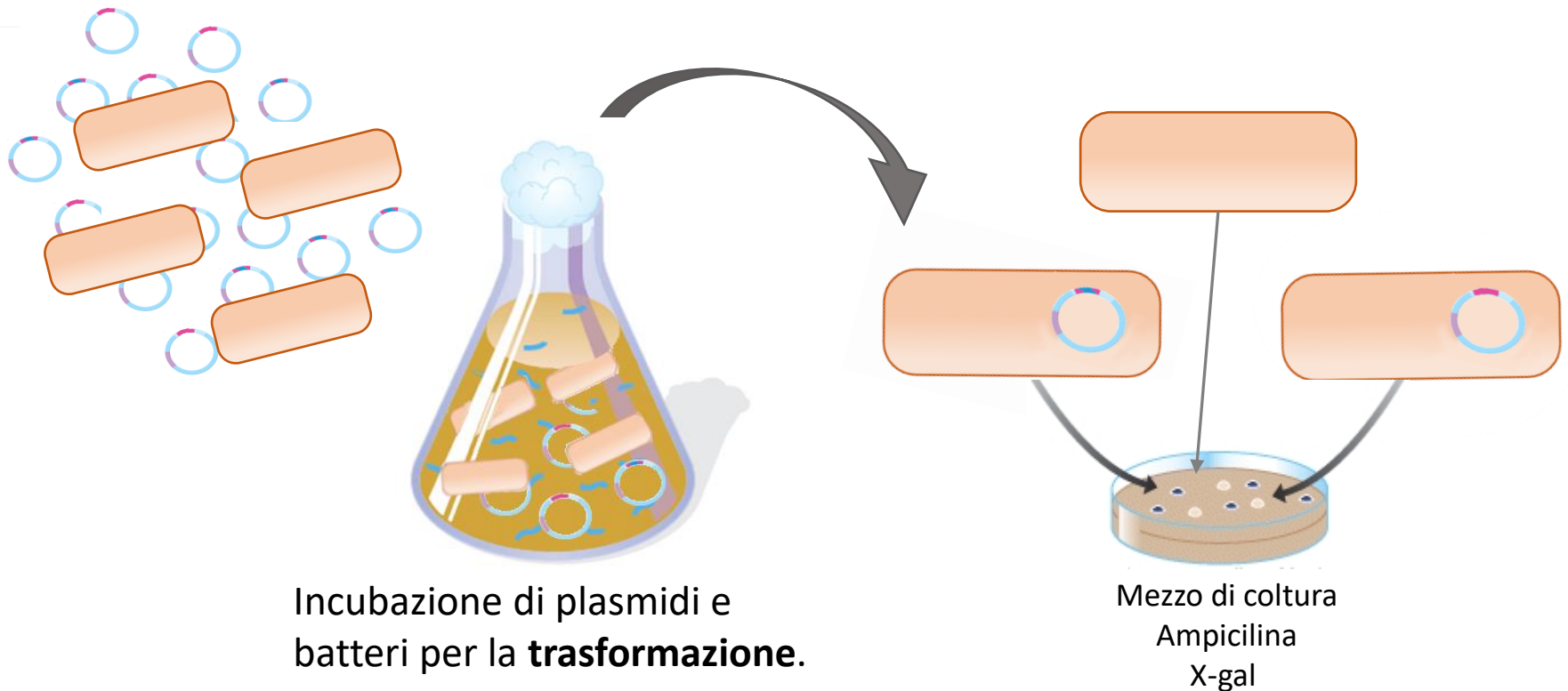
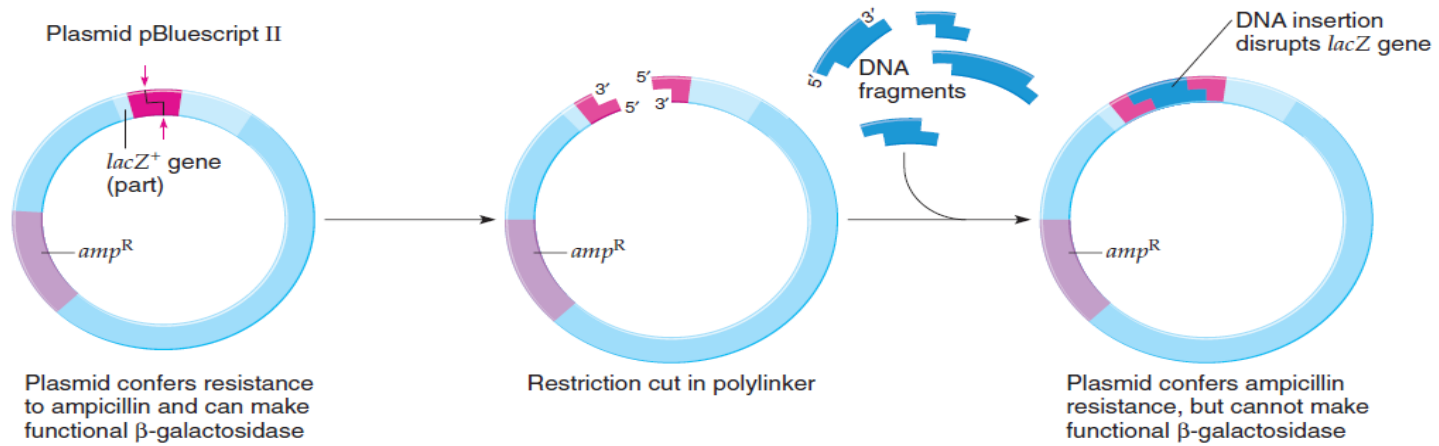
- **Plasmidi.** Nei vettori di clonaggio plasmidici derivati di *E.coli* possono essere clonati frammenti di DNA fino a 15 kb.
- **Batteriofagi**
- **Cosmidi:** vettori che presentano caratteristiche sia dei vettori plasmidici sia dei batteriofagi. Possono contenere inserti di DNA di 40-45 kb.
- **Cromosomi artificiali.** Sono i vettori che possono accogliere i frammenti più grandi di DNA

Vettori di clonaggio plasmidici: derivati da plasmidi circolari trovati in natura nei batteri. Questi vettori vengono modificati in laboratorio (“ingegnerizzati”) perché abbiano le seguenti caratteristiche:

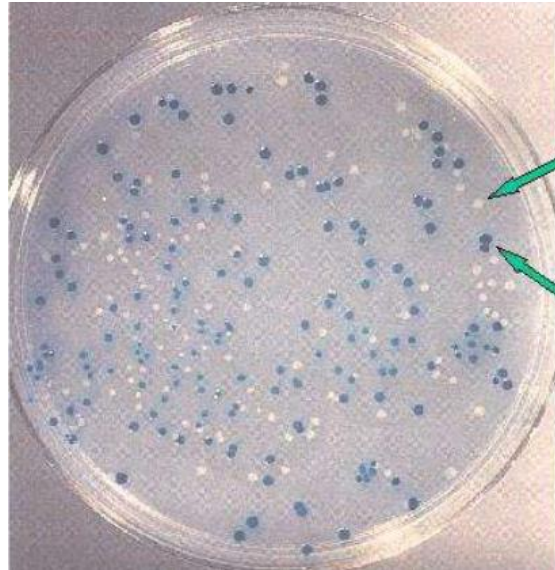


Vettore di clonaggio plasmidico pBluescript II.

Figure 8.5 Insertion of a piece of DNA into the plasmid cloning vector pBluescript II to produce a recombinant DNA molecule.



Incubazione di plasmidi e batteri per la **trasformazione**.

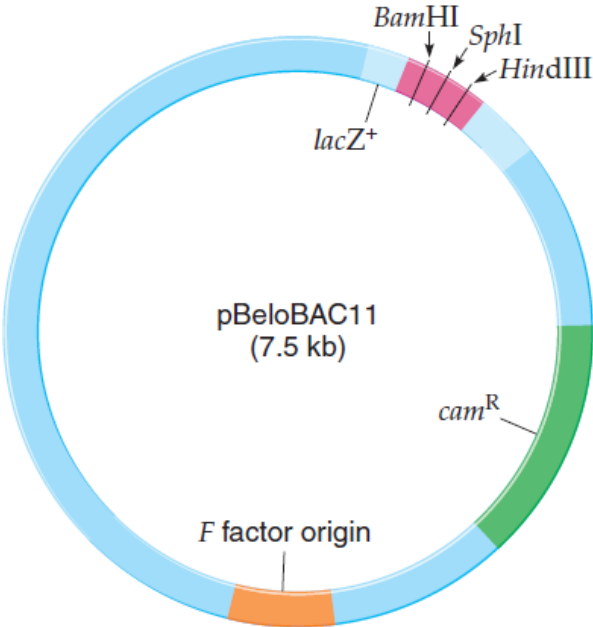


Colony with insert

Colony with no insert

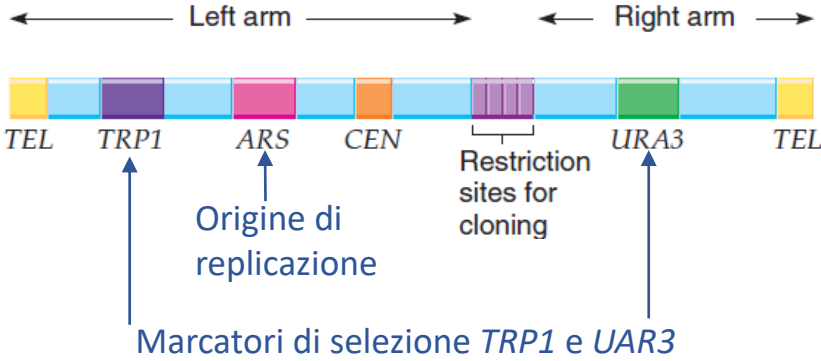
Cromosomi artificiali: I cromosomi artificiali sono vettori di clonaggio che possono accogliere frammenti di DNA molto grandi, formando molecole di DNA ricombinante simili a piccoli cromosomi.

a) A bacterial artificial chromosome (BAC) vector



cam^R = Chloramphenicol resistance gene
lacZ⁺ = Part of β -galactosidase gene

b) A yeast artificial chromosome (YAC) vector



I vettori YAC sono quelli che possono contenere gli inserti più lunghi di DNA (fino a 2 Mb)
 Questi cromosomi artificiali non vengono replicati nei batteri, ma nelle cellule di lievito.
 Il YAC è un vettore lineale

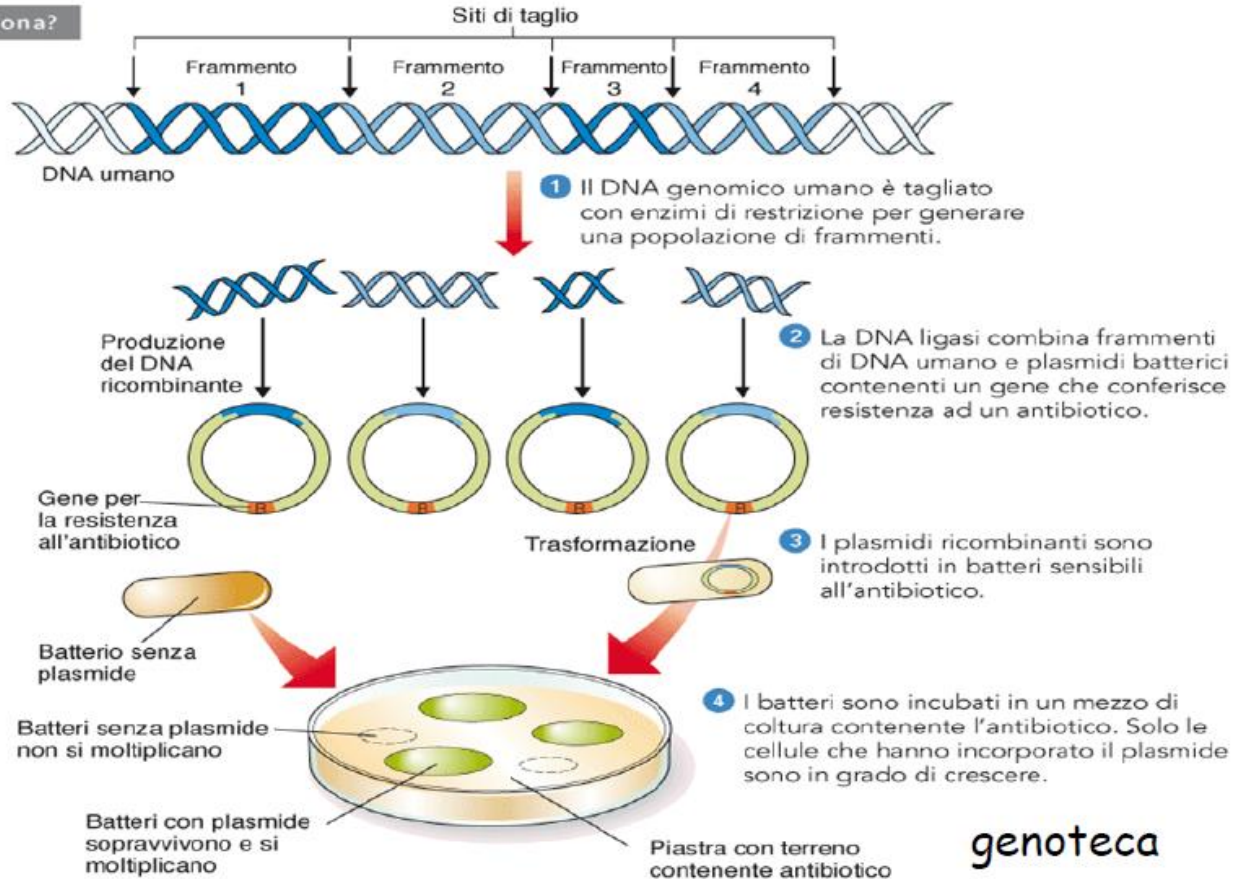
I BAC possono contenere inserti fino a 300 kb.

Costruzione di banche genomiche (anche chiamata libreria genomica o genoteca)

Perché si usa?

Una libreria genomica è una raccolta di tutti i frammenti di DNA del genoma di un organismo; una libreria cromosomica è una raccolta di tutti i frammenti di DNA di un cromosoma specifico. Ciascun frammento di DNA viene "conservato" in un clone di cellule batteriche ricombinanti. Gli scienziati usano librerie genomiche e cromosomiche per isolare e studiare geni.

Come funziona?



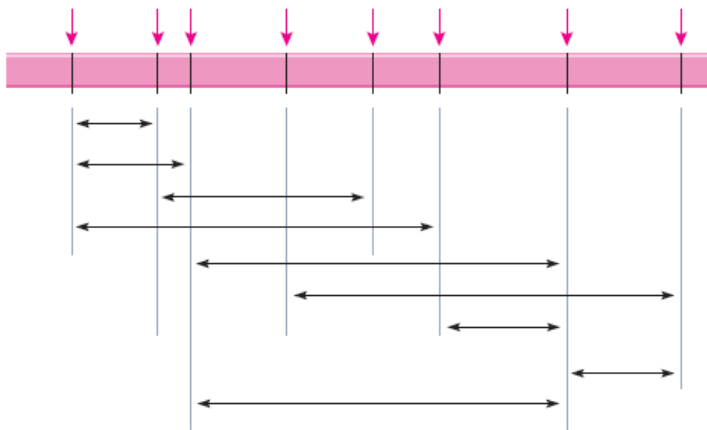
Banche cromosomiche: In una genoteca genomica, la collezione di frammenti di DNA clonati derivati da un singolo cromosoma viene chiamata banca o libreria cromosomica. Nel genoma implica la costruzione di 24 genoteche cromosomiche diverse (una per ciascuno degli 22 autosomi più i cromosomi sessuali X e Y).

Digestione genomica:

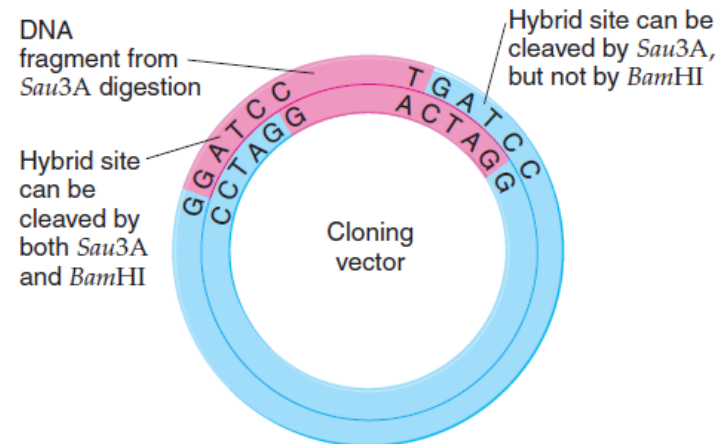
- 1) frammentare il DNA genomico in modo di ottenere frammenti di dimensioni simili e adeguata alla capacità del vettore di clonaggio.
- 2) per poter ricostruire l'ordine di queste frammenti nel genoma, occorre che in essi ci siano delle sequenze sovrapponibili.

Due fattori sono fondamentali a questo scopo: la scelta degli enzimi di restrizione e il tempo di digestione (*digestione parziale* del genoma).

a) Partial digestion of DNA by a restriction enzyme (for example *Sau3A*) generates a series of overlapping fragments, each with identical 5' GATC sticky ends

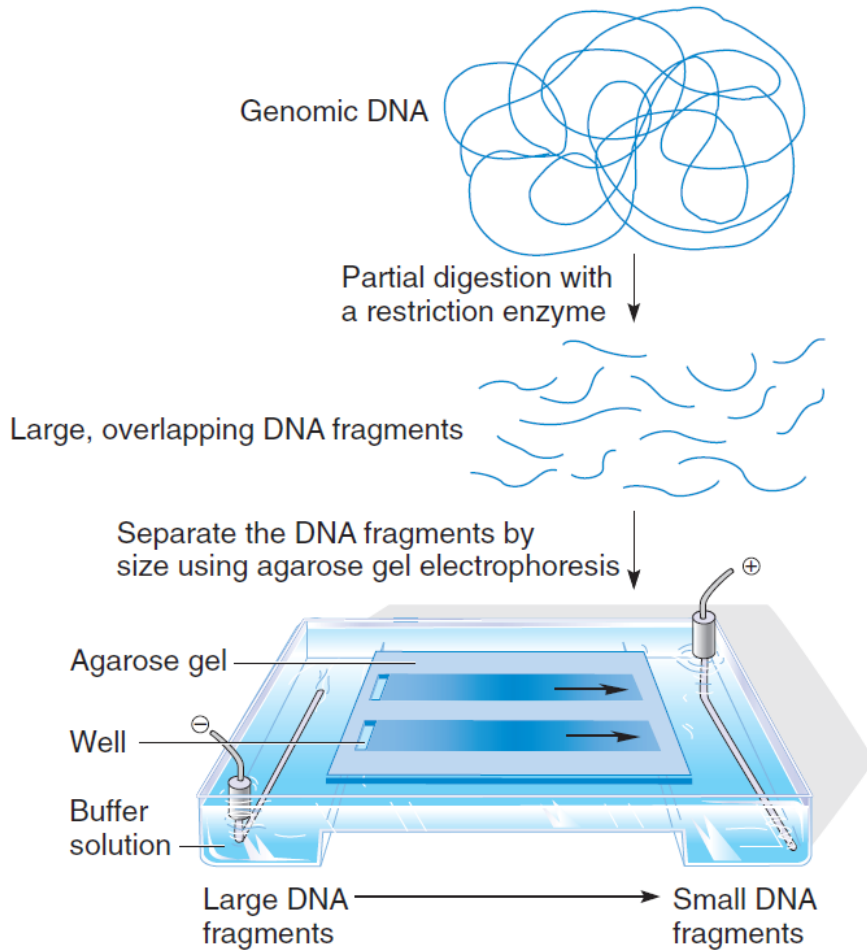


b) Resulting fragments may be inserted into *Bam*HI site of cloning vector

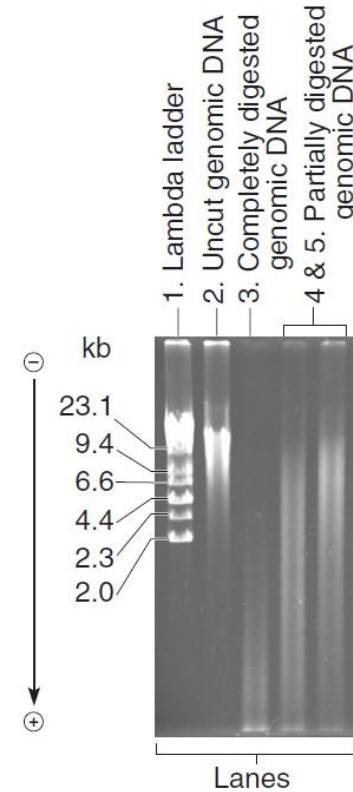


La dimensione dei frammenti risultanti può visualizzarsi per **elettroforesi dei frammenti di DNA su gel di agarosio**.

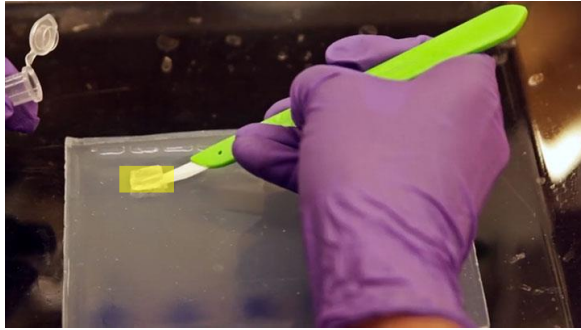
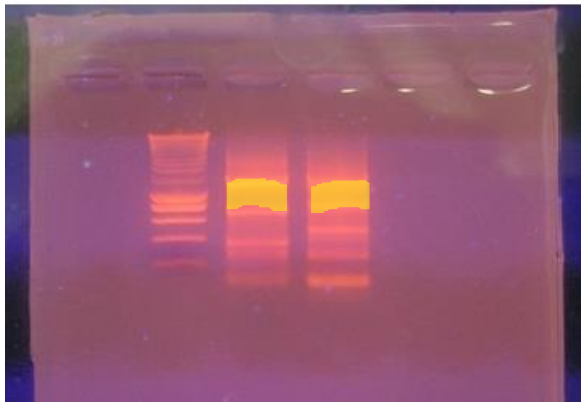
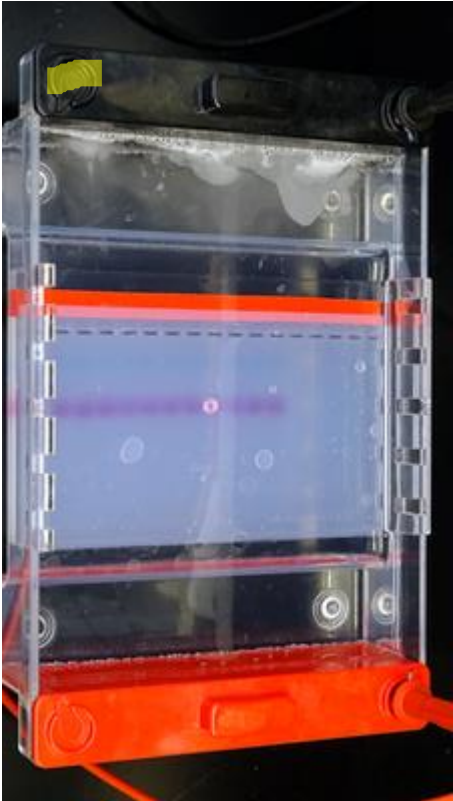
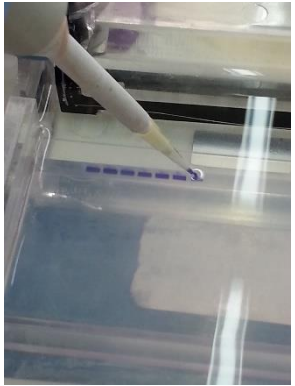
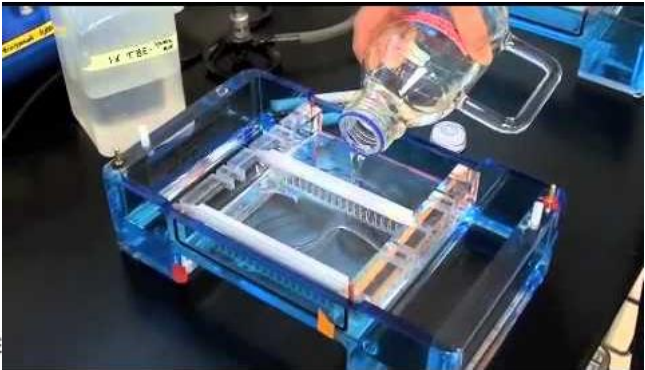
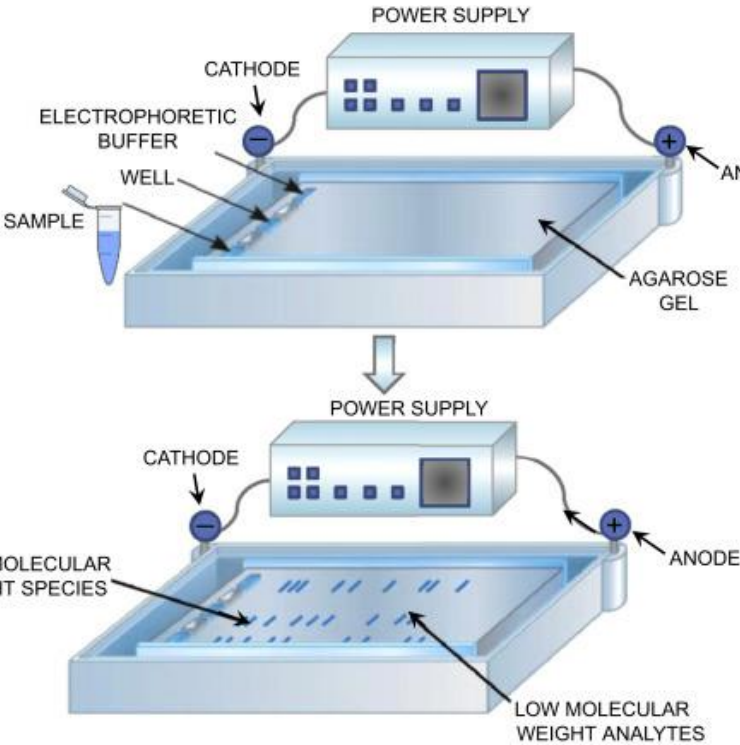
a) **Partial restriction digestion of genomic DNA.**



b) **Agarose gel electrophoresis analysis of genomic DNA partially digested with restriction enzyme.**



Agarose gel electrophoresis



Sequenziamento del DNA

Al giorno di oggi esistono due metodi principali di sequenziamento del DNA:

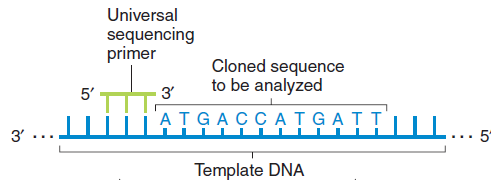
- Il metodo classico o **metodo di Sanger**, sviluppato da Fred Sanger e basato nel sequenziamento di DNA mediante didesossinucleotidi.
- E una tecnologia più moderna, di sviluppo recente, la **Next Generation Sequencing**.

Tutti i due si basano nel processo nella reazione di sintesi di una nuova catena di DNA complementare alla catena stampo.

I. Sequenziamento Sanger o sequenziamento mediante didesossinucleotidi.

Elementi della reazione di sequenziamento:

- Il DNA isolato dal clone batterico (**DNA stampo** da sequenziare).
- **Primer**: Un oligonucleotide a singolo filamento che serve per l'inizio della reazione di sintesi di DNA. La sequenza del primer è complementare alla sequenza del plasmido adiacente all'inserto di DNA.

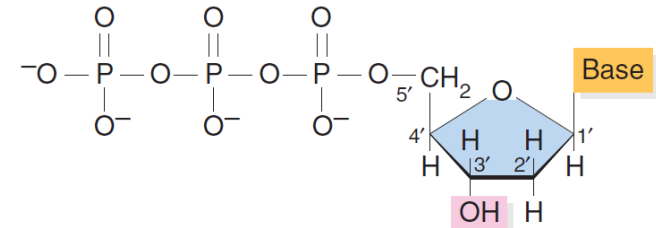


- Una enzima **DNA polimerasi** che porta avanti la reazione di sintesi del DNA
- Quattro **desossinucleotidi** precursori (dNTP, ossia dATP, dTTP, dCTP e dGTP), e una quantità inferiore (~1:100) di nucleotidi modificati, chiamati **didesossinucleotidi** (ddNTP, ossia ddATP, ddTTP, ddCTP e ddGTP).

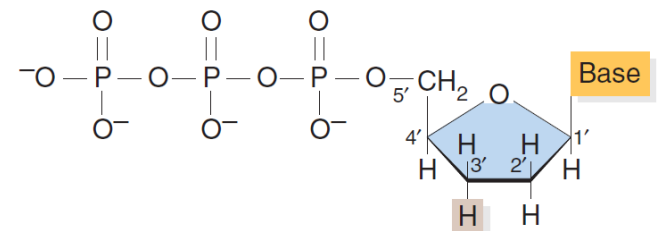


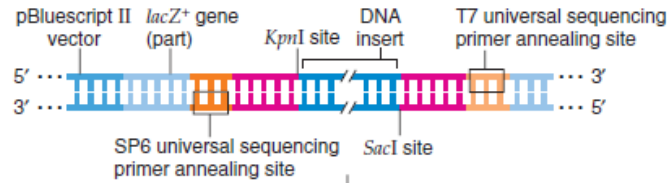
Sequenziatore automatico
ABI PRISM 3100 Genetic Analyzer

a) Deoxynucleotide (dNTP) DNA precursor



b) Dideoxynucleotide (ddNTP) DNA precursor





Denature to single strands

Unione dei primer:



Reazione di sintesi e sequenziamento:

Figure 8.11a Dideoxy sequencing.

a)

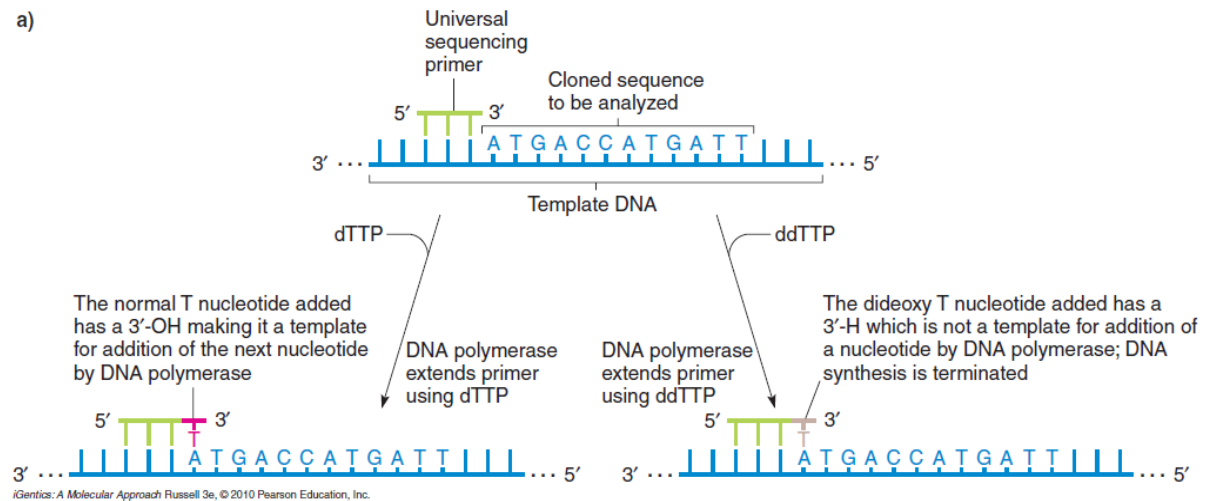
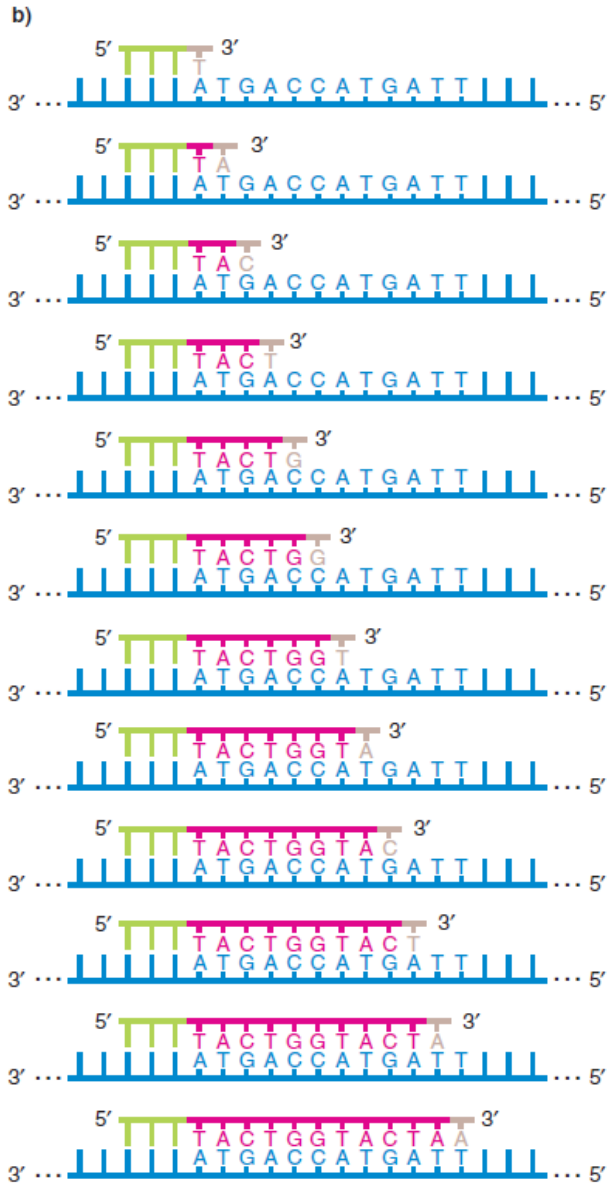
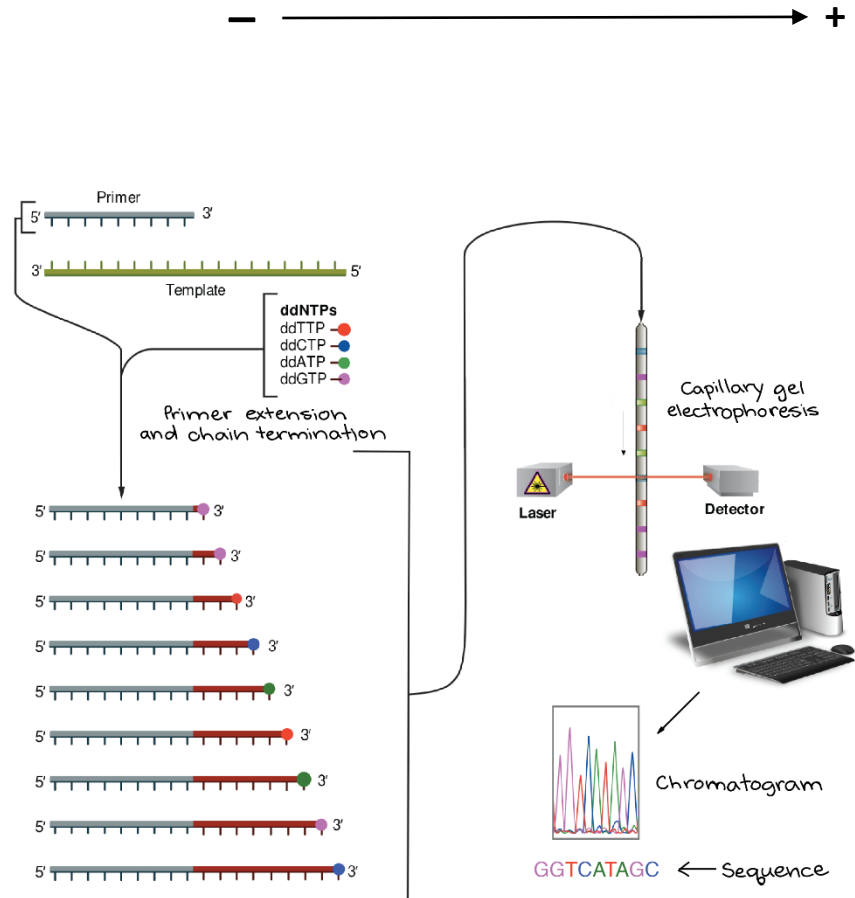
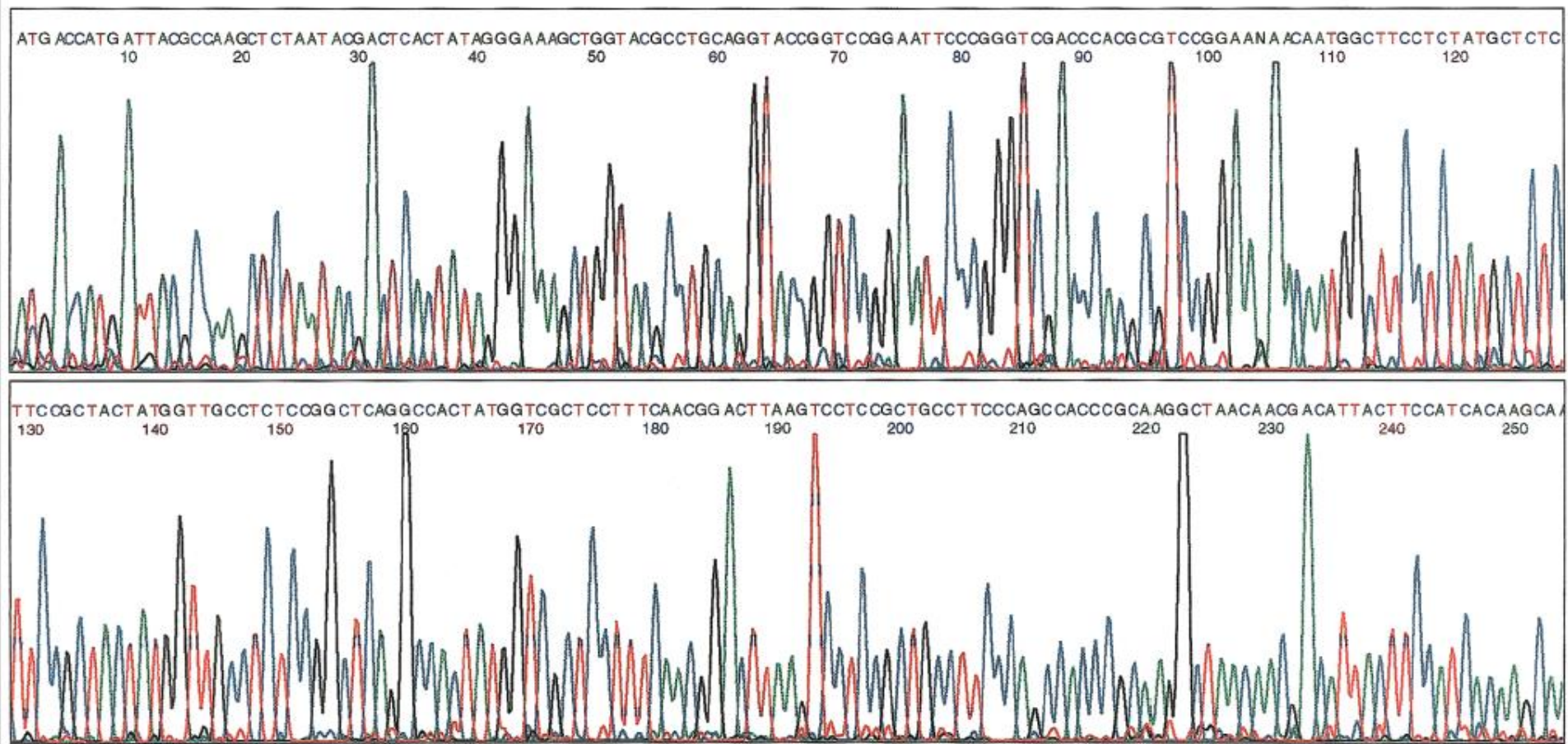


Figure 8.11b Dideoxy sequencing.



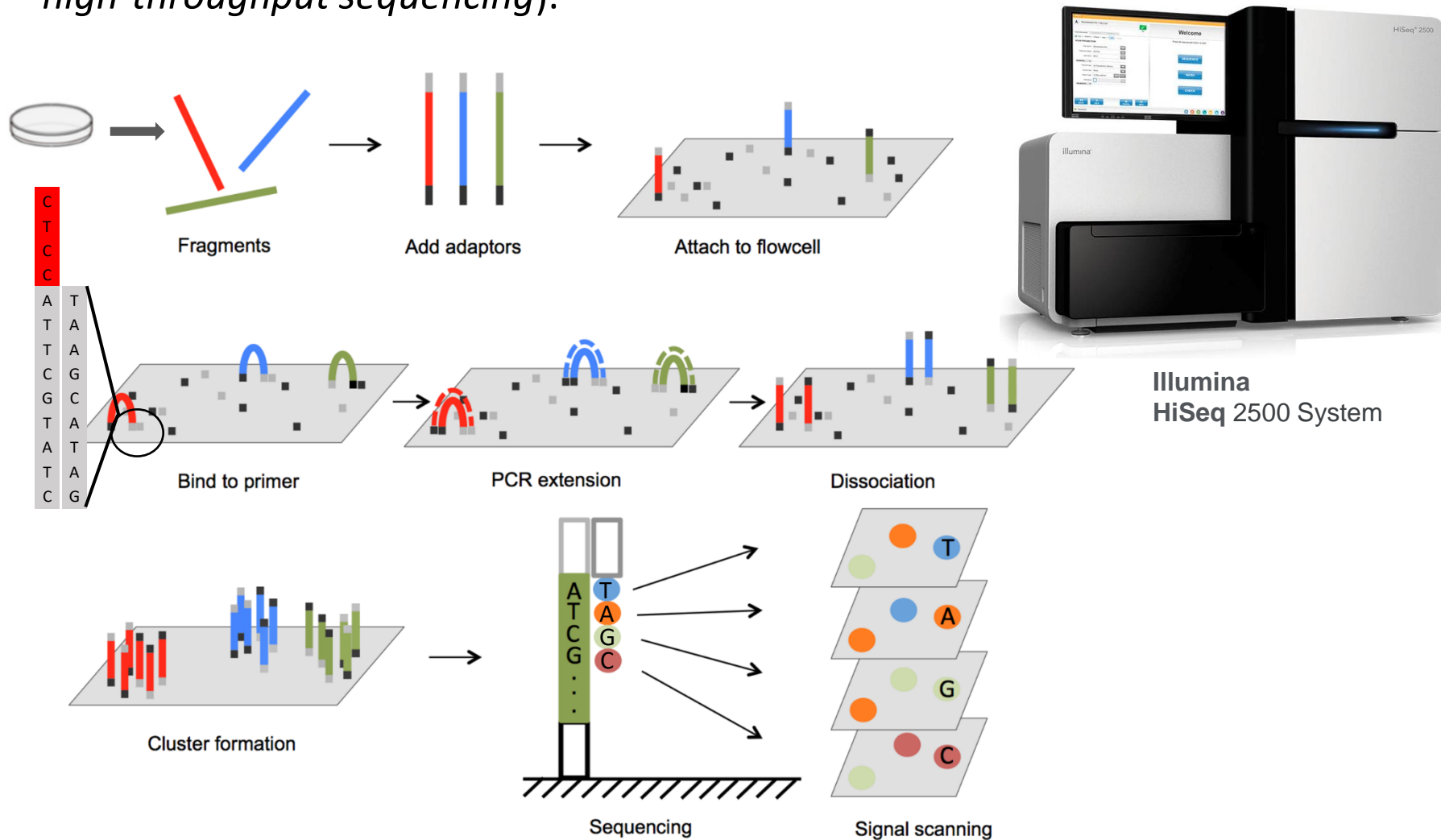
In una reazione di sequenziamento, ci sono milioni di coppie di primer/stampo identiche che subiscono la stessa reazione di allungamento. Quindi, statisticamente, alcune reazioni si fermano al nucleotide 1, altre al 2, altre al 3, etc. Nel complesso, si sintetizzeranno molte molecole di DNA di nuova sintesi con un numero elevato di frammenti che terminano in ogni posizione.





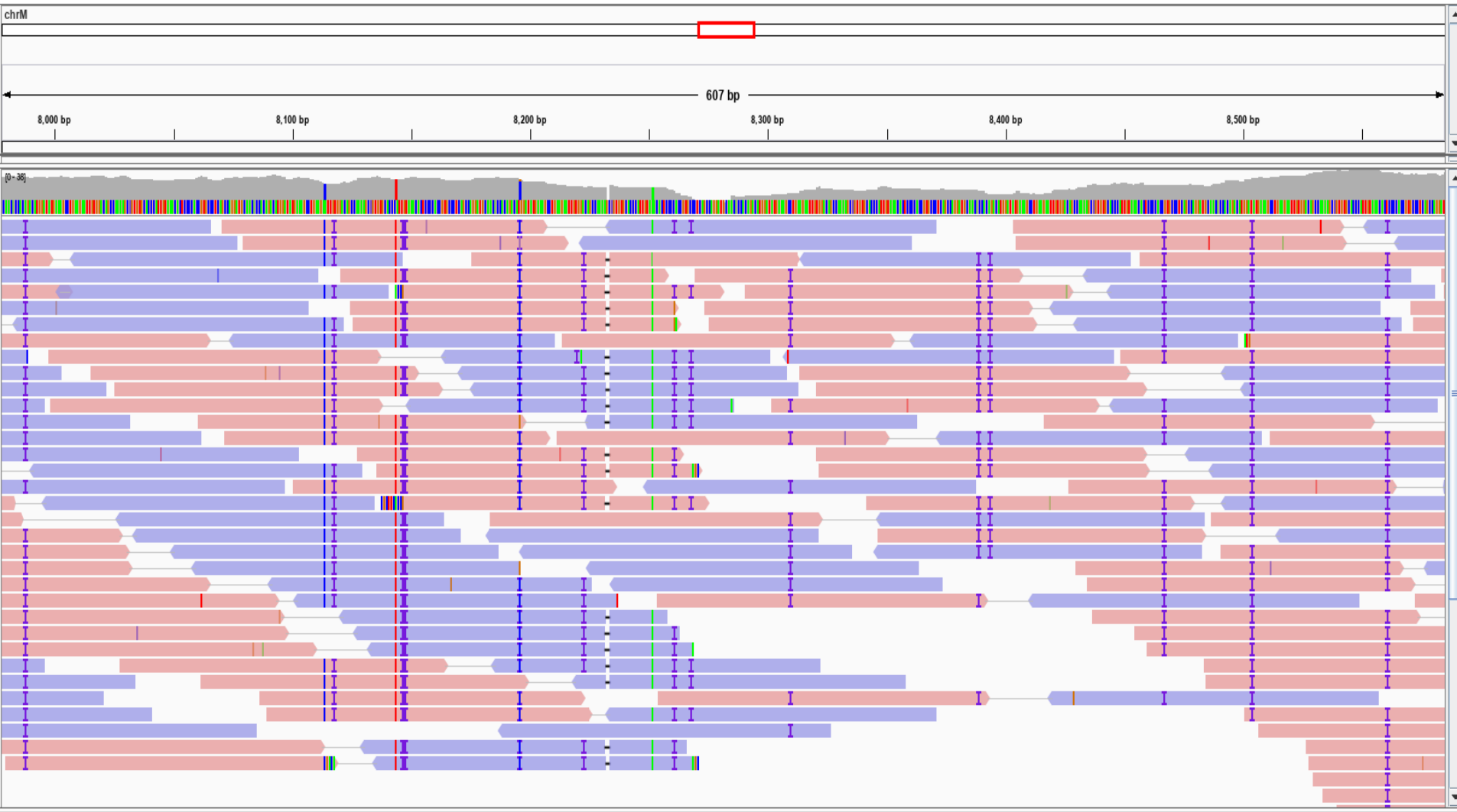
Le sequenze di DNA risultanti del sequenziamento si chiamano *reads*

II. Next Generation Sequencing (NGS, tecnologie di nuova generazione o *high-throughput sequencing*).



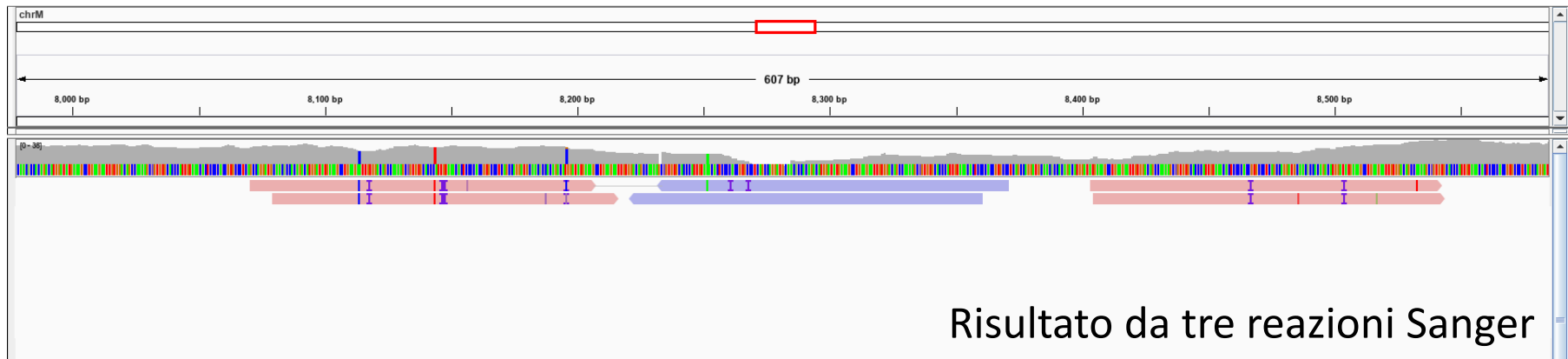
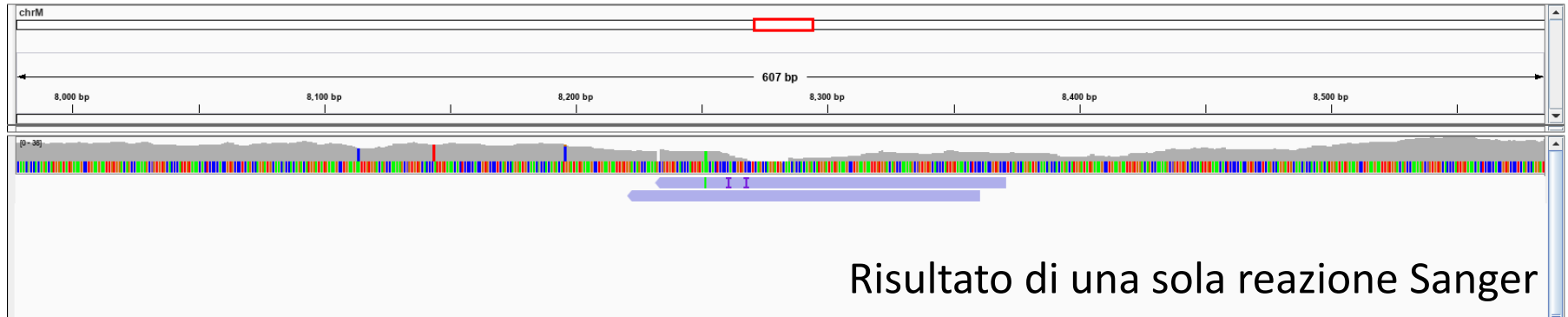
Ogni reazione di sequenziamento NGS produce del ordine di centinaia di *reads* per ogni frammento di DNA stampo.

NGS: ogni reazione permette il sequenziamento simultaneo di migliaia di frammenti di DNA e, la produzione di centinaia di *reads* per ogni frammento di DNA stampo.



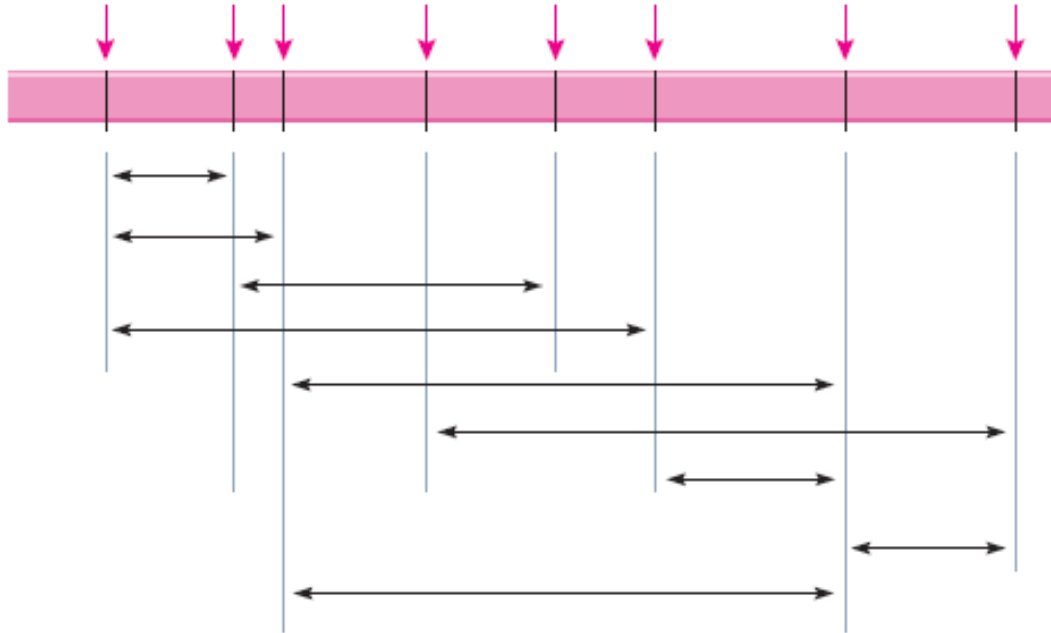
Risultato di una sola reazione NGS

Sanger: ogni reazione permette il sequenziamento di un unico frammenti di DNA e produce un massimo di due *reads* (forwards 5'-3' e reverse 3'-5').

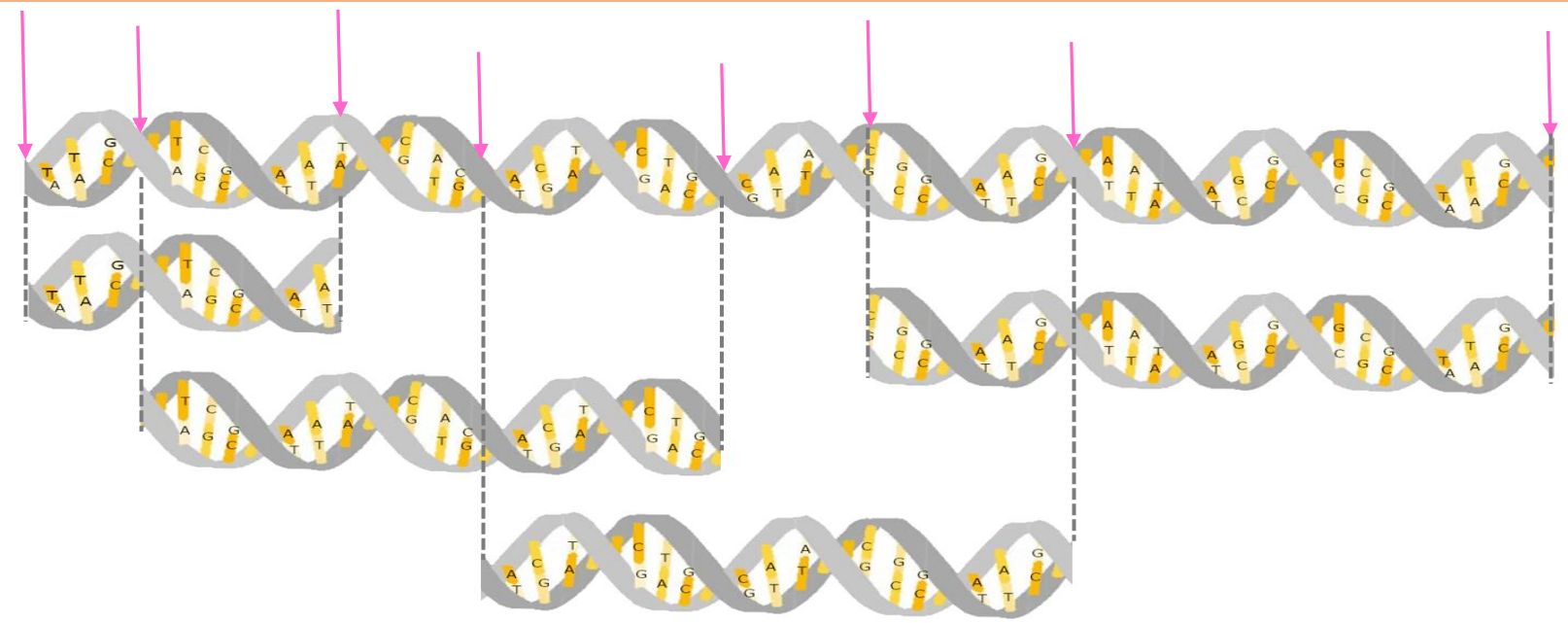


Ricostruzione della sequenza completa di DNA

La ricostruzione della sequenza di DNA originale si basa nell'assemblaggio dei *reads* in base alle sequenze che si sovrappongono.



Risultato della digestione enzimatica: Frammenti sovrapposti



Risultato del sequenziamento: Reads sovrapposti

5'-AACAGCTT-3'

5'-AGCTTAGTGTGAGAC-3'

5'-TGAGACGTTGCCTTC-3'

5'-CCTTCTTATCCCGCAAC-3'

Assemblaggio della sequenza di DNA:

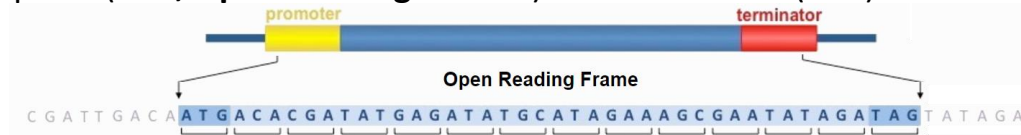
5'-AACAGCTTAGTGTGAGACGTTGCCTTCTTATCCCGCAAC-3'

Annotazione delle sequenze genomiche: Identificazione dei geni

Due metodi complementari:

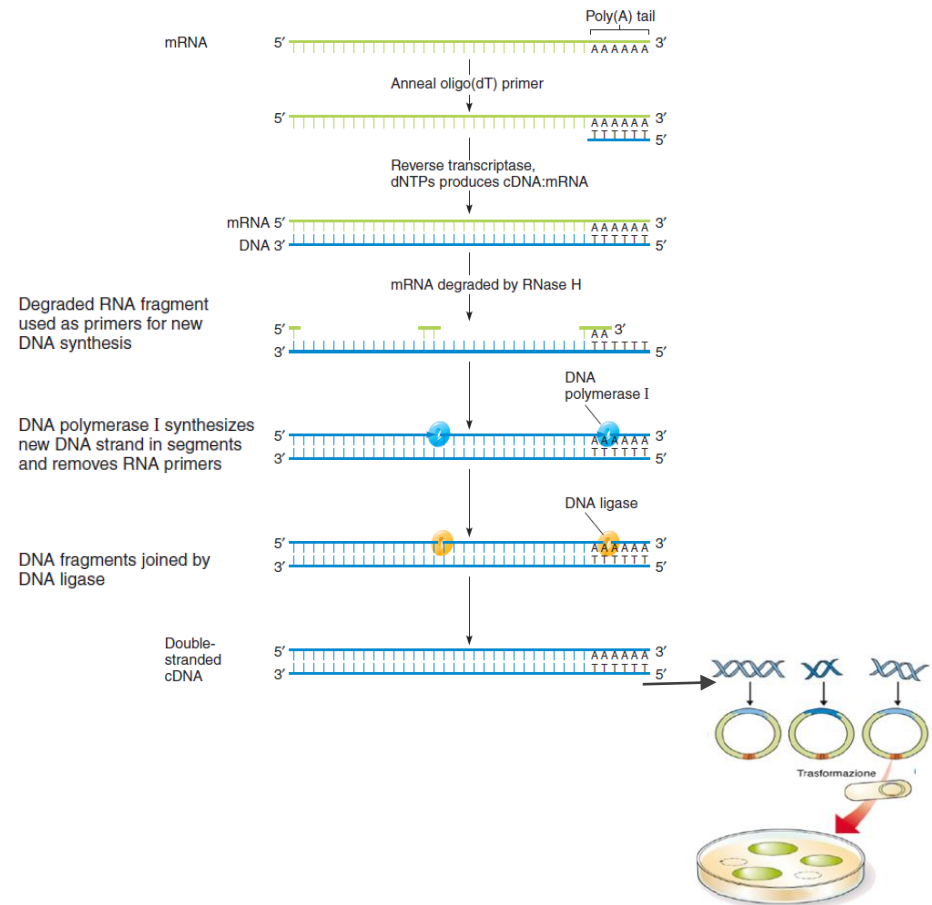
I. Metodi computazionali

Identificazione computazionale delle fasi di letture aperte (**ORF, Open Reading Frames**): Codoni d'inizio (ATG) in fase (separati da multipli di tre nucleotidi) e codoni di stop (TAG, TAA o TGA).

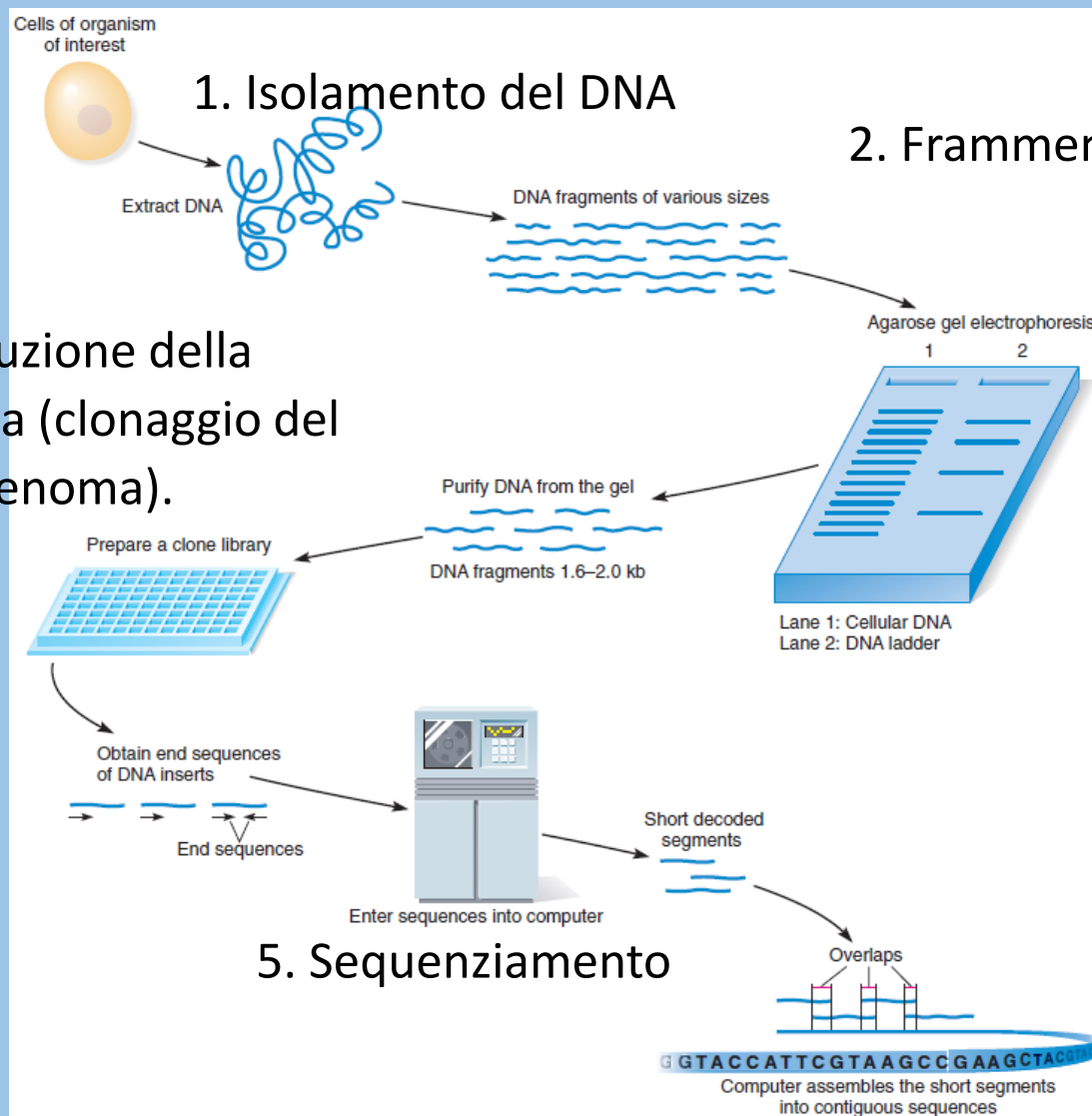


II. Clonaggio e sequenziamento di cDNA.

1. Isolamento del **RNAm**.
2. Le molecole di **RNAm vengono usate come stampo per la sintesi di molecole complementari di DNA (cDNA)**. Enzime coinvolte: trascrittasi inversa (RT, Reverse Transcriptase), RNasi-H (un tipo di ribonucleasi), DNA polimerasi e DNA ligasi. Si ottiene una molecola di cDNA a doppia elica che è una copia dell'mRNA di partenza.
3. Inserzione del cDNA (1-5 kb) in vettori di clonaggio.
4. Trasformazione delle cellule ospiti di E.coli con i vettori ricombinanti.
5. **Sequenziamento del cDNA** clonato.
6. I *reads* derivati di questo cDNA vengono poi confrontati con le sequenze genomiche per l'identificazione dei geni.



Promemoria: Sequenziamento genomico



4. Costruzione della genoteca (clonaggio del intero genoma).

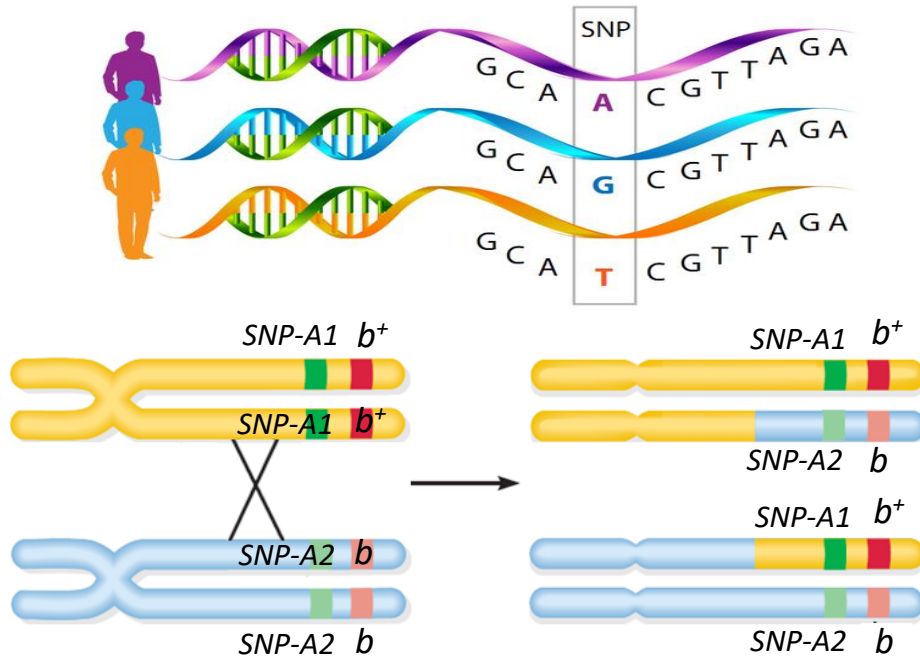
3. Selezione dei frammenti di DNA di dimensione adeguata per il clonaggio

6. Assemblaggio delle sequenze di DNA

Questo approccio di sequenziamento genomico si conosce come **sequenziamento casuale diretto** (*whole-genome shotgun*)

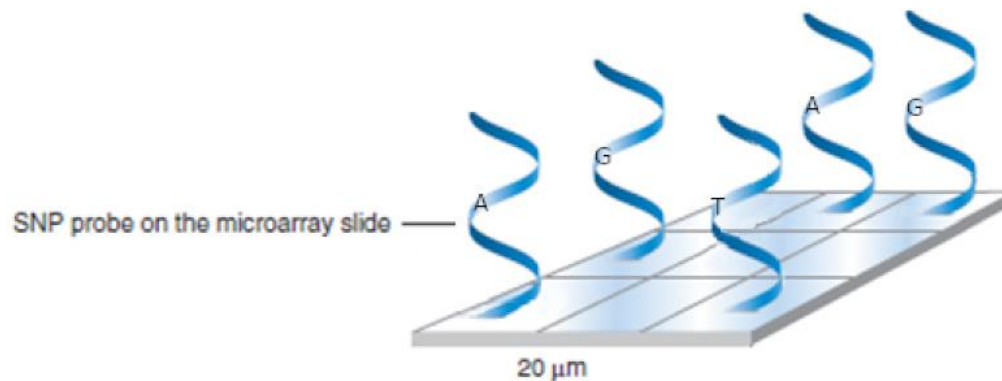
Sequenziamento in **microarray** (o **chip di DNA**)

Sequenziamento di marcatori **SNP** (*Single nucleotide polymorphism*).



Aplotipi: Sequenze di varianti alleliche che si trovano vicini in una piccola regione di un cromosoma. Nell'esempio, la combinazione di alleli *SNP-A1 b+* costituisce un aplotipo.

Microarray di sequenziamento:



Sequenziamento in microarray (o chip di DNA)

Cells of organism of interest

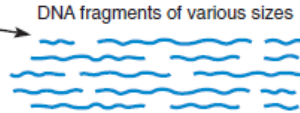
1. Isolamento del DNA



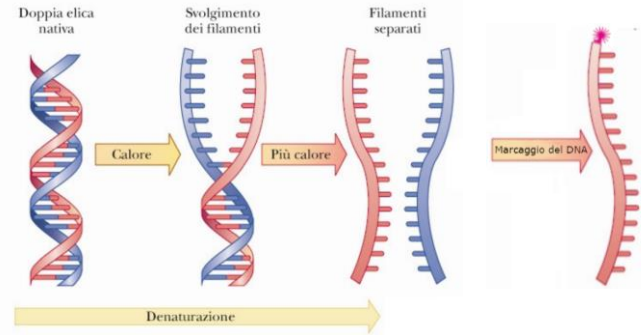
Extract DNA



2. Frammentazione del DNA



3. Denaturalizzazione del DNA

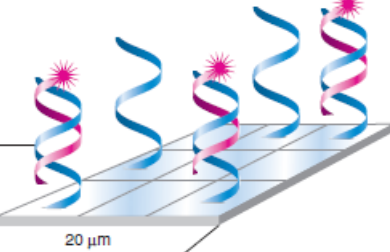


4. Costruzione della genoteca (clonaggio del intero genoma).

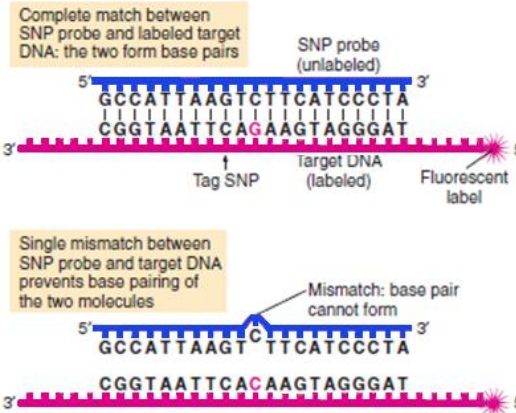
b)

Diagram of part of the hybridized probe cell

Individual's labeled genomic DNA containing a SNP allele binds to the SNP probe on the microarray slide if the match is perfect



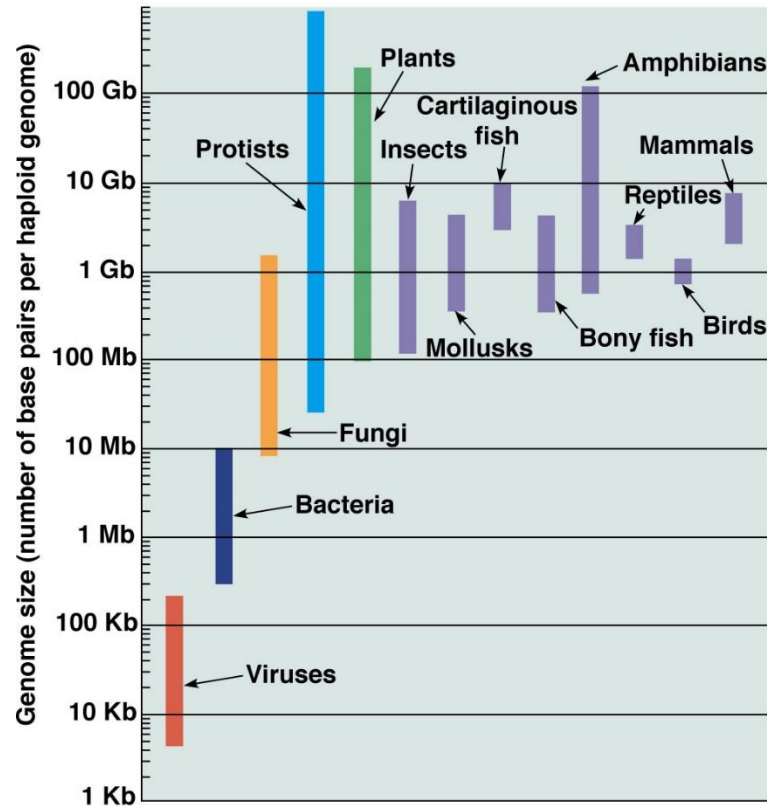
Individual's labeled genomic DNA containing a SNP allele does not bind to the SNP probe on the microarray slide if the two sequences are not perfectly complementary



4. Ibridazione con le sonde SNP (in base al principio di complementarità delle basi, si forma una doppia elica ibrida tra la sonda SNP e il DNA denaturato e marcato.

Paradosso del valore C

Scarsa correlazione fra dimensioni del genoma e complessità dell'organismo

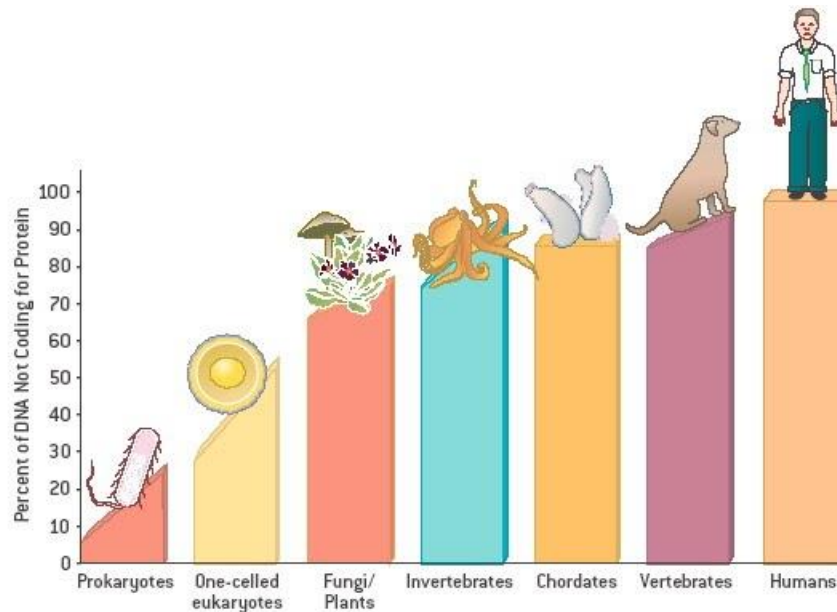


Densità genica: numero di geni per data lunghezza del DNA

Common name or class	Scientific name	Genome size (Mb)	Number of genes	Gene density (genes/Mb)
Eukaryotes				
Baker's yeast	<i>Saccharomyces cerevisiae</i>	12	6,241	480
Nematode	<i>Caenorhabditis elegans</i>	97	18,424	190
Cruciferous	<i>Arabidopsis thaliana</i>	125	25,498	204
Fruit fly	<i>Drosophila melanogaster</i>	180	13,601	75
Pufferfish	<i>Fugu rubripes</i>	400	35,000	100
Sea urchin	<i>Strongylocentrotus purpuratus</i>	900	27,350	30
Sea urchin	<i>Strongylocentrotus purpuratus</i>	900	27,350	30
Human	<i>Homo sapiens</i>	3,400	35,000	10
Archaea				
Crenarchaeota	<i>Aeropyrum pernix</i>	155	1,522	981
Euryarchaeota	<i>Methanococcus jannaschii</i>	166	1,715	1033
Euryarchaeota	<i>Archaeoglobus</i>	218	2,420	1110
Bacteria				
Gram positive	<i>Mycoplama genitalium</i>	0.58	479	831
Proteobacteria	<i>Buchnera sp. APS</i>	0.64	564	881
Gram negative	<i>Haemophilus influenzae</i>	1.8	1,727	959
Cyanobacteria	<i>Synechocystis sp.</i>	3.6	3,168	880
Gram positive	<i>Bacillus subtilis</i>	4.2	4,100	976
Proteobacteria	<i>Escherichia coli</i>	4.6	4,288	932

Table 1. Genome size, gene number and gene density.

C'è, invece, una correlazione fra frazione del genoma non codificante e complessità dell'organismo



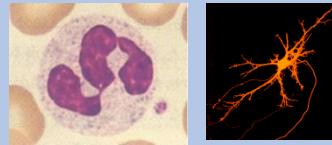
NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

Table of Human Genome Statistics

Topic	Statistic
Total size of the genome:	approximately 3,200,000,000 bp (male, haploid)
Percentage of adenine (A) in the genome:	54%
Percentage of cytosine (C) in the genome:	38%
Percentage of bases not yet determined:	6%
Highest gene-dense chromosome:	chromosome 19 with 23 genes per 1,000,000 bp*
Least gene-dense chromosomes:	chromosome 13 and Y with 5 genes per 1,000,000 bp*
Percentage of DNA spanned by genes:	between 25% and 38%
Percentage of exons:	1.1 to 1.4%
Percentage of introns:	24% to 37%
Percentage of intergenic DNA:	74% to 64%
The average size of a gene:	27,000 bp*
The longest gene:	dystrophin (a muscle protein) with 2,400,000 bp*
Average length of an intron:	3,300 bp*
Most common length of an intron:	87 bp*
Occurrence rate of SNPs:	roughly 1 per 1,500 bp
Occurrence rate of genes:	about 12 per 1,000,000 bp

Ma quanto siamo diversi? Quante basi del DNA lo sono?

Due cellule dello stesso individuo



0/1000

Due gemelli identici



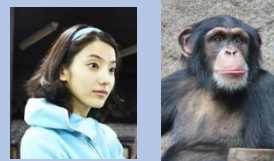
0/1000

Due di noi a caso



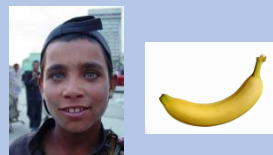
1/1000

Uno di noi e uno scimpanzè



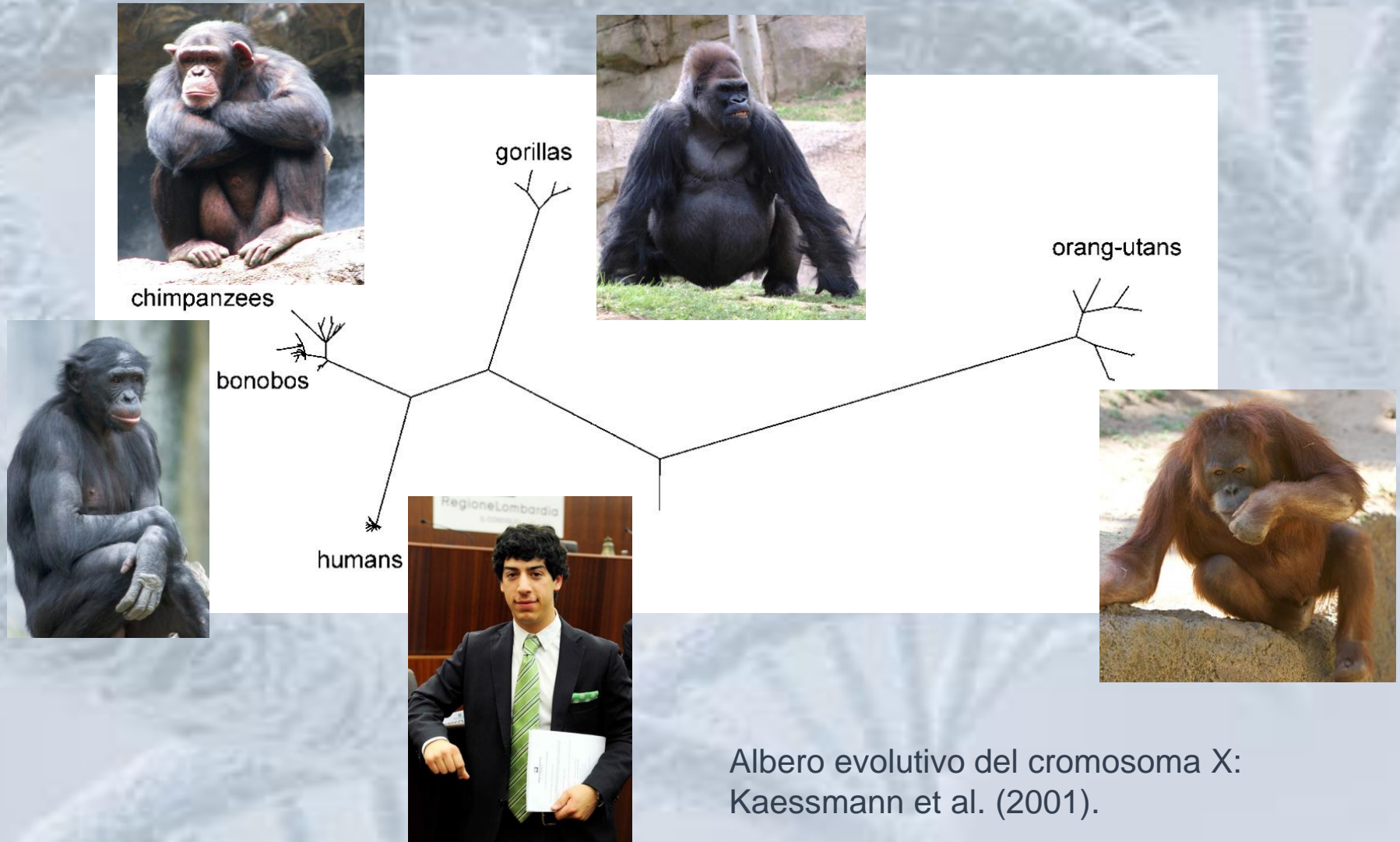
10-30/1000

Uno di noi e una banana



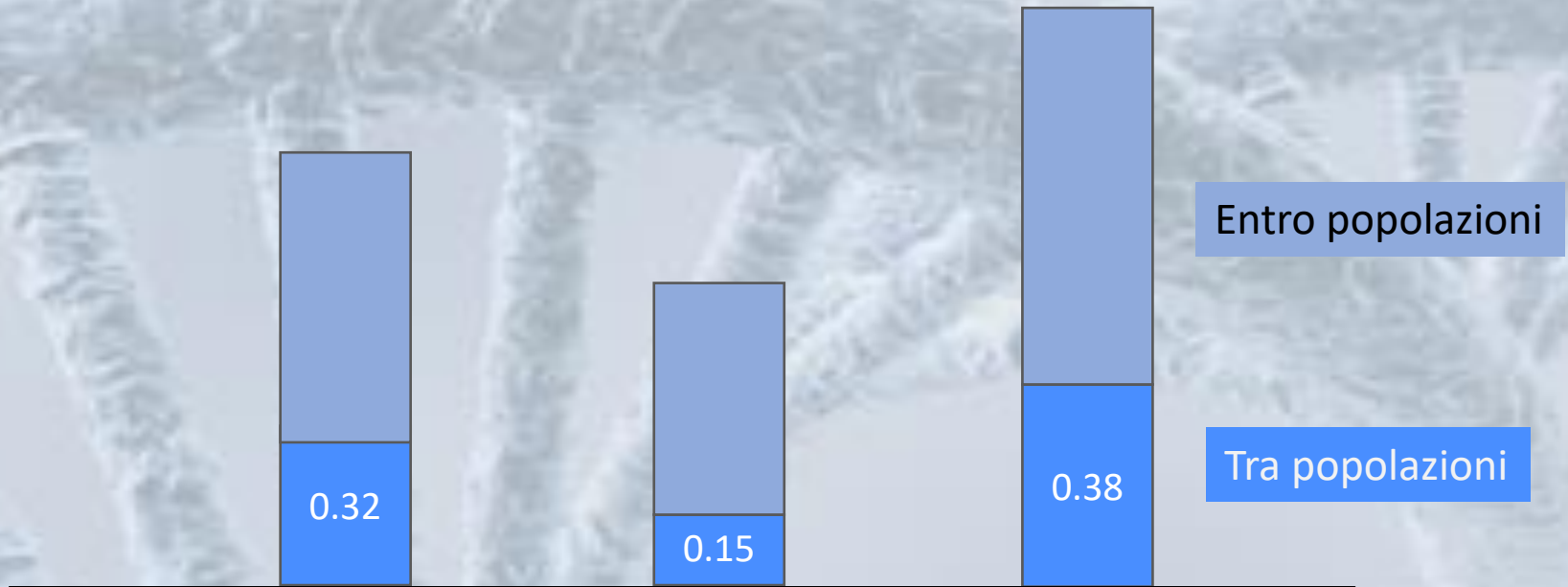
750/1000

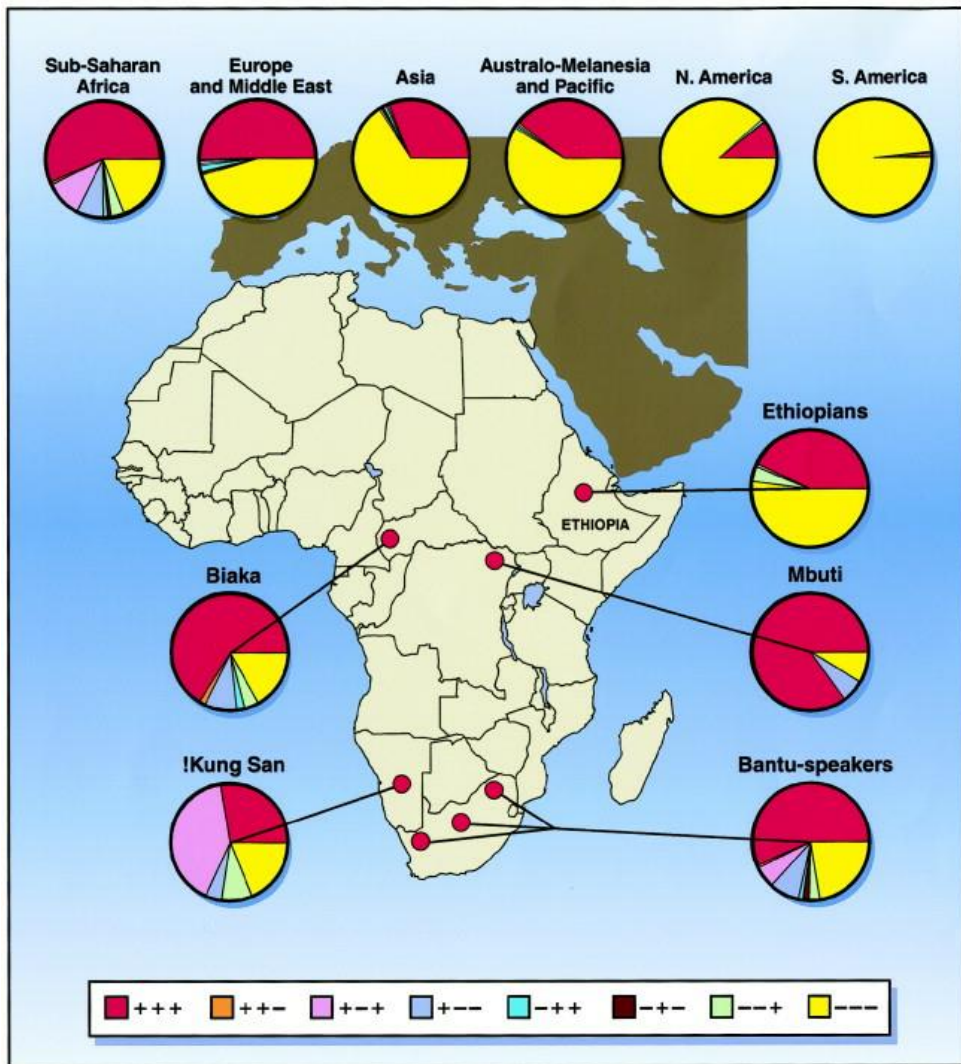
Le differenze genetiche nella nostra specie sono le più basse di tutti i primati



Albero evolutivo del cromosoma X:
Kaessmann et al. (2001).

Le differenze genetiche fra popolazioni umane sono le più basse fra tutti i primati





Africa is special:
Genetic diversity in
all continents is
often a subset of
African genetic
variation

Tishkoff et al. (1998)

A pointillist view of human evolution and variation: 1

Early Homo sapiens sapiens
in Africa

150,000 to 100,000 BP



A pointillist view of human evolution and variation: 2



Homo sapiens sapiens
colonizing south west Asia

~100,000 BP

A pointillist view of human evolution and variation: 3

